# Investigating the genetic control of complex traits

**Richard Fordjour Oppong**

Doctor of Philosophy (PhD) – The University of Edinburgh – 2018

# Contents

# Declaration

I declare that I have wholly undertaken the study reported herein and that except portions where references have been duly cited, this thesis is the outcome of my research. No part of this thesis has been submitted to another institution for any other degree or professional qualification.

……..

Richard Fordjour Oppong

*To my parents:*

**Nana Yaa Emo** and **Kwadwo Fordjour**

# Acknowledgements

Undertaking this research has been a wonderful experience for me from start to finish and it wouldn't have been so without the brilliant support I received from a lot of people. First is my supervisor, Sara Knott, to whom I am enormously grateful for granting me the opportunity to have a go at this research. I thank her and the rest of my supervisory team, Chris Haley and Pau Navarro, for their undying patience for allowing me to learn and develop at my own pace while my PhD took shape, and for the many hours of advice and support they have given me throughout my PhD.

I would also extend my gratitude to present and past members of the Haley group, especially Charley Xia, Yanni Zeng, Carmen Amador and Masoud Shirali for their useful discussions and comments and for sharing pieces of computer code with me. Many thanks to Jarrod Hadfield, my thesis committee member, and Ian White from whom I benefited a lot in terms of gaining technical statistical knowledge. I also benefited from the many discussions I had with John Woolliams and I thank him for his effective ways of explaining difficult concepts to me.

I thank the Darwin Trust of Edinburgh for generously funding my PhD for four years. I thank the staff of Edinburgh Parallel Computing Centre at the University of Edinburgh for keeping Eddie oiled and working to run my analysis.

I thank my family for their unflinching support and encouragement throughout the course of my PhD. Finally, I thank all the many friends I made whilst studying in Edinburgh.

“Nea onnim no sua a, ohunu”

– Asante proverb

# Lay summary

For any trait of interest to the geneticist, such as disease phenotypes and economically important agricultural phenotypes, there are differences in the measurements taken for any set of individuals randomly chosen from the population. These differences or variation may be attributed to two sources, the environment of the individuals and their genes. There is particular interest to estimate the contribution of the genes to the observed variation in measured phenotypes because such values influence actions aimed at human disease management and improving agricultural yield. For instance, a large genetic contribution to a phenotype of interest will mean efforts should be made to identify genes involved and that further efforts should be made to know the mechanisms by which these genes influence the phenotype. But in instances where a low genetic contribution to phenotypic variation is estimated, it will be prudent to look into the environment of the individuals for answers. There is therefore the need to develop effective ways of estimating the genetic contribution to the observed phenotype differences because the values inform important decisions in genetics research. Biased or inaccurate estimates will undermine efforts aimed at disentangling the genetic underpinning of phenotypic variation and consequently disease management or improvement of agricultural yield. This thesis examines the performance of some of the existing statistical methods used and develops novel statistical methods to highlight how violations of model assumptions impact the estimates of genetic contribution to the phenotypic variation. It was found that the inappropriate use of methods affects the estimates obtained and that there are benefits in using methods which capture important features driving phenotype variation.

# Abstract

One aspect of the effort to disentangle the genetic component of complex traits is the mapping of genetic loci that are associated with variations in complex traits. The model used to assess these genetic associations is vital because any inaccurate or biased estimate obtained will undermine the effort to unravel the genetic component of complex traits. This project, therefore, explores in detail the performance of existing analytical methods and develops novel statistical methods to bring to the fore how violations of the model assumptions impact model estimates. Using a linear mixed effects model, a genome-wide association analysis of eight urine phenotypes was performed on 2,934 individuals, where it was shown that violations of the model assumption of normality of residuals impact heritability estimates and subsequent genetic associations. The issue of normality was explored further in a simulation study that used a Bayesian mixture model that assumed that the effect SNPs of complex traits are drawn from a mixture of four normal distributions instead of one. The Bayesian mixture model was applied to the urine phenotypes and it was shown that the effect SNPs are constituted from more than one normal distribution and over 99% of genotyped SNPs were sampled to have no effect on the urine traits. The urine traits were also shown to have a polygenic architecture with much of the additive genetic variance being explained by SNPs with small to moderate effects. Departing from SNP based approaches, I present a novel analytical approach that utilises a relationship matrix that is based on natural haplotype blocks defined by recombination boundaries in the genome. This method was developed on the premise that haplotypes provide a better strategy for capturing the true genomic relationship amongst individuals in the presence of rare variants and thus provide real benefit over SNPs in recovering much of the hidden heritability of complex traits and in the identification of novel gene variants. The method was implemented on simulated data and was explored in detail. The results from the simulation showed that the haplotype approach complemented existing GWAS analytical approaches by capturing regions in the genome contributing to the phenotypic variation that existing GWAS methods fail to capture. It was further demonstrated that there are real benefits to be gained from this approach by applying it to real data from circa 20,000 individuals from the Generation Scotland: Scottish Family Health Study. Height and major depressive disorder were analysed, and novel genomic regions were identified for both traits. In conclusion, this thesis shows that inappropriate use of analytical models can impact results which may have consequences on conclusions we draw from genetic association studies. Also, the thesis shows the benefits of implementing models that capture important features of the underlying architecture driving the variation in complex traits. Lastly, the thesis also demonstrates that haplotype methods can complement conventional SNP-based methods in the bid to understand the genetic control of complex traits.

# List of figures

# List of Tables

“In God we trust; all others must bring data.”

— W. Edwards Deming

# Chapter 1

## 1 Introduction

The prospects of the Human Genome Project (HGP) (Lander et al., 2001) were exciting in many respects. Perhaps, the single most exciting amongst them was being able to deliver health benefits from DNA sequence information, by providing explanations to the genetic basis of medically relevant phenotypes (Collins et al., 2003).

To deliver this promise, investigators began exploring the human genome through linkage disequilibrium mapping which revealed patterns in the distribution of DNA sequence variants (Gibbs et al., 2003). This essentially led to the cataloguing of the most common source of variation in the genome, Single Nucleotide Polymorphisms (SNPs), to try and unlock the genetic basis of observed traits variation. Nearly two decades now after the completion of the HGP and with development of high throughput genotyping technologies that genotype these SNPs, numerous studies designed to find disease risk SNPs across the whole genome have tried to explain the causes of and responses to common chronic diseases (that affect

a large proportion of the population) by estimating their association to these SNPs (Burton et al., 2007; McCarthy et al., 2008).

These whole-genome studies have tremendously improved our understanding of the genetic component of the phenotypic variation, although much of the genetic variation in most traits remains to be characterized and explained. It should be noted also that these genetic association studies as we know them today are a culmination of ground-breaking work spanning three centuries back to the 19th Century. Right from the days of Darwin, the study of trait inheritance – later known as genetics – has progressed steadily to the 21st-century science we know today and with it the development of our understanding of the genetic basis of observed traits, more importantly, diseases.

## 1.1 Our evolving understanding of the genetics of traits; from Darwin to genome-wide association studies (GWAS)

Ultimately, Mendel's work was crucial to the general acceptance of Darwin's theory of evolution by natural selection. Although this was not immediately obvious when Mendel was rediscovered in the early 20th century (Hill, 1984). Mendel had discovered that traits are inherited in a particulate or discrete manner (Moore, 2001). This upon rediscovery, rather naively offered a disparate view from Darwin's view of blending inheritance of traits (Moore, 2001), in which progeny are seen as intermediate between parents (Darwin, 1859). Trait variation, as observed by Darwin and argued by the biometricians (Hill, 1984), was a continuous progression that blended into each other gradually in an unbroken manner which makes it almost impossible to pinpoint where one becomes the next. This was a clear departure from

the apparently large and discrete variation in traits as put forward by Mendel. This sparked a debate about the basis of inherited traits.

The continuous versus discrete debate and how they explain evolution by natural selection dominated trait inheritance discourse in the very early days of the field until laboratory experiments on artificial selection offered some resolution (Hill, 1984). Notable were Morgan's experiments on the patterns of inheritance of characters in Drosophila (Morgan and Bridges, 1916) which showed that there was a link between Mendel's discrete inheritance and Darwin's gradualist and continuous inheritance. Morgan and his colleagues showed that the aggregate effect of many different loci behaving in a discrete Mendelian manner could produce the gradual and continuous distribution of traits as observed by Darwin to explain evolution (Kennedy, 2001). This effectively put to bed a hotly contested debate and allowed focus perhaps to be shifted onto the mathematical quantification of the trait variation.

Ronald Fisher, in his paper presented to the Royal Society of Edinburgh, showed that the observed variation in a trait and correlation between relatives could be used to partition the observed variation (phenotype) into the unobserved genetic (genotype) and environment components (Fisher, 1918). He further partitioned the genetic component of trait variation into the additive, dominant and epistatic components. And it was Fisher who originally proposed a way to directly estimate the contribution of genes to the variation observed in traits, which became known as the heritability.

Fisher's works together with works such as the estimation of inbreeding effects by Wright and further work by Haldane offered a lasting theoretical resolution to the initial debate and helped set the stage for the mathematically intensive field of quantitative genetics in the 1930s (Hill, 1984). Wright defined the inbreeding coefficient $\boldsymbol{f}$ – the correlation between gametes produced by the two parents. This was later estimated as the probability that two alleles at any locus in an individual are Identical by Descent (IBD) from the common ancestor of the two parents (Hill, 1984), thus making the concept of inbreeding useful in the estimation of relationships between individuals and essentially the estimation of the genetic variance (Powell et al., 2010).

With the phenotypic variation mathematically defined, research focus shifted to using genetic markers to quantify the genetic component of the variation or heritability. Understanding the genetic variation, both in humans and in other organisms, presents huge medical and agricultural benefits and may also offer some useful explanations of the evolutionary process. For instance, in humans, heritability estimates showed that fitness-related traits like fertility had low estimates whilst traits like height had very high estimates, offering a view of how selection acts on alleles with respect to their role in fitness.

Initial studies on heritability estimation were done by regressing offspring trait values against the average trait values of parents (Falconer, 1960). Later studies that incorporated molecular information were largely made possible by the introduction of the technique of protein electrophoresis into population genetics

(Harris and Hopkinson, 1976). Consequently, variations at the protein level were assayed for in several organisms.

The advent of DNA based techniques offered researchers the opportunity to look directly into the genetic material to measure genetic variation. Restriction Fragment Length Polymorphisms (Botstein et al., 1980) and Southern blotting techniques (Southern, 1975) were among the first DNA based techniques developed. Later the polymerase chain reaction (PCR) (Mullis et al., 1986) was developed and with it an array of techniques which enabled large-scale analysis of genetic variation through microsatellite markers.

Then came along genomics and its hopes of offering an all-encompassing approach to study genetic inheritance which would provide biologists with the opportunity to study traits (mostly diseases) in organisms to an extent not previously thought possible. The successes of the human genome project (Lander et al., 2001), made this genomics era a reality (Collins et al., 2003).

Eventually, the International HapMap Consortium, in 2003, employing an approach which had become widely used to study genetic variation, fully sequenced subsamples of different human populations to identify set of polymorphisms such as Copy Number Variations (CNVs) and SNPs (Gibbs et al., 2003). The most common form of variation identified by the consortium were SNPs (Gibbs et al., 2003; Frazer et al., 2007) and thus SNP information was used to develop genotyping arrays which would be used in studies to genotype these SNPs in much larger human samples (Burton et al., 2007; Frazer et al., 2007). Association testing of genotyped SNPs with

phenotypic variation would become known as a Genome-Wide Association Study (GWAS) (Burton et al., 2007; McCarthy et al., 2008). There have since been more than 3,000 GWAS published with more than 59,000 unique SNP – trait associations reported (MacArthur et al., 2017).

## 1.2 The genetic traits

All traits of interest to the geneticist fall under two broad categories. These are Mendelian traits and complex traits. Mendelian traits or disorders are monogenic which means that a single gene is sufficient to cause phenotype. They segregate in families, but can also be sporadic, e.g. new mutations arising de novo in singletons. They can occur in several modes which are grouped according to whether they are dominant or recessive, and autosomal or sex-linked (Chial, 2008). Conditions such as haemophilia, cystic fibrosis, sickle-cell anaemia, and polycystic kidney disease are examples of Mendelian disorders.

Mendelian disorders are usually individually rare (Antonarakis and Beckmann, 2006) and the variants that cause them are highly penetrant (Bodmer and Bonilla, 2008), which means a greater number of carriers of the disease-causing alleles develop the associated disease conditions. These alleles are usually kept at very low frequencies in populations by selection because of their large and deleterious effect on fitness (Bodmer and Bonilla, 2008). The range and frequencies of alleles controlling these Mendelian disorders differ among populations (Björkegren et al., 2015).

Multiple genetic and environmental factors affect most common diseases like diabetes, hypertension, coronary artery disease, chronic kidney disease and biometric traits like height, weight, and body mass index. And as a result, these traits are termed complex traits. The genetic variants that confer risk of development of these traits are many and may individually explain only a small fraction of the genetic variation (Burton et al., 2007). These variants have low to moderate penetrance (Bodmer and Bonilla, 2008) and could be highly influenced by the environment (Burton et al., 2007).

## 1.3 The variation in traits

For any trait, like height, for example, there are differences in the measurements taken for any set of individuals sampled from a population. These differences in the trait values may be the results of two main influences. Influences attributed to the genes of the individuals and influences collectively attributed to the environment of the individuals.

The genetic contribution to the differences in the trait can be ascribed primarily to the randomness that is introduced by Mendelian segregation during gamete formation and to random mating and random fertilisation. These create differences in the measured trait values based on what set of alleles an individual inherited because different alleles may affect a trait differently. The alleles can also introduce systematic differences by virtue of an individual being a male or female: for instance, for height, on average, males will be taller than females in the population.

The environmental factors that generate differences in the observed trait values are numerous. For starters, the process of trait measurement itself can generate differences. That is, the measuring instrument may lack finer graduations and thus introduce rounding errors in the measured trait values. Different people taking repeated measurements of the trait may generate differences. Other factors influencing the life history of sampled individuals such as access to essential nutrition (in our example trait, height, lack of essential nutrients can result in stunting) or individual's lifestyle such as whether an individual drinks or smokes (this influences measures of cardiovascular traits), may generate differences in measured trait values.

So, for every individual sampled in the population, the measured trait value will deviate from an expected value or population mean by some value attributable to the factors described above and others. This deviation is termed the error – where error here does not necessarily refer to some mistake made but rather to the difference between an observed trait value and the expected value.

Statistically, if we start from the premise that the study individuals are randomly sampled from the study population then we can assume that the observed deviations around the expected value (i.e. the errors) are random numbers coming from some statistical distribution. We can, therefore, model a vector of trait values, $\boldsymbol{y}$, with length $n$ linearly as

$$y = X\beta + e \tag{1.1}$$

where $\boldsymbol{\beta}$ is a $k \times 1$ vector of $k$ model parameters, $\boldsymbol{X}$ is an $n \times k$ matrix called the design matrix and it contains columns that assign the observed trait values to elements in the parameter vector $\boldsymbol{\beta}$. The expected trait value given the model, i.e. the expectation of $\boldsymbol{y}$, $E[\boldsymbol{y}] = \boldsymbol{X\beta}$, and $\boldsymbol{e}$ is a vector of errors.

The standard linear model used in genetics studies assumes that the errors (deviations) are sampled from a normal distribution. Thus, two parameters are of importance. The first parameter is the mean around which the observed trait values cluster. The second is a parameter that describes the dispersion of the observed values around this mean. In describing this dispersion parameter for say height, it is intuitive to see that individuals shorter than expected will deviate negatively from the mean whiles those taller than expected will have a positive deviation. Summing the positive and negative deviations may cancel out when calculating the average deviation from expectation for a sample of individuals. Therefore, the average of the square of these individual deviations from the expectation is calculated instead. This is the variance of the trait or trait variation. And given the linear model (equation 1.1), the vector of traits $\boldsymbol{y}$ is assumed to be normally distributed $\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \boldsymbol{I}\sigma_e^2)$, $\boldsymbol{I}$ is an $n \times n$ identity matrix and $\sigma_e^2$ is the variance of the errors (trait variance).

The trait variance is of primary interest to geneticists and animal breeders and it is to be estimated. Most importantly, we like to quantify how much of this variance is due to the environment and how much is down to the genetics of study individuals. If a trait is under sufficient genetic control (i.e. much of the trait variation is explained by variation in genes) then it will be prudent to implement policies and interventions that target the genes. If on the other hand much or all the variation is

due to the environment, then interventions that target the environment could be implemented.

## 1.4 The genetic component of the trait variation

The trait variation, as already mentioned, can have a genetic component and a part that is attributed to the environment. The genetic component of the variation could be further dissected into three parts, the additive component – which is the aggregate sum of the effects of all gene variants that directly affect a trait, the dominant component – which is the effect of the interaction between the alleles of a gene at a given gene locus, and the epistatic component – which is the effect of the interaction between different gene variants at different gene loci.

Heritability is the term used to describe the genetic component of the trait variation. It can be defined in the broad sense which will encompass contributions from all the three components of the genetic variation. It can also be defined in the narrow sense which includes a contribution from only the additive genetic component.

Genetically, there are two categories of traits, Mendelian and complex traits, and these two categories have very different genetic architectures which make the approaches to studying and estimating their heritability also different.

### 1.4.1 The study of Mendelian traits

The heritability of Mendelian traits has traditionally been estimated using pedigree-based studies of families with individuals affected by these conditions (Chial, 2008). The study of the inheritance process of Mendelian traits provides useful

insights into the genetic mechanism underlying all forms of genetic diseases (Antonarakis and Beckmann, 2006).

The successes of familial and linkage studies in helping to understand rare Mendelian variants inspired the studies of complex traits and thus have been argued to be given more prominence in genetic studies (Antonarakis and Beckmann, 2006). But in a world of limited resources and funds, studying complex traits is rather more appealing. This is because such traits affect so much of the population and thus offer a far greater benefit in terms of the reach of research output and commercial gains, and thus there is a bigger focus on the study of complex traits currently. The field receives more funding from research councils and pharmaceutical companies.

### 1.4.2 The study of complex traits

Twins were used initially to estimate the genetic variation or heritability of complex traits in humans. The twin methods were grounded on the assumption that twins share the same environment (Kendler et al., 1993), such as the same prenatal environment and the same home environment. Thus, identical (monozygotic) and fraternal (dizygotic) twins should be equally correlated to the environmental factors that influence the trait of interest. Therefore, if there is excess similarity observed in monozygotic twins over dizygotic twins for a trait under study, then the trait is affected by genetic factors. The genetic contribution to the trait variation can then be estimated because monozygotic twins completely share their genetic information while dizygotic twin share just half of it on average and thus will have half of the twin pair correlation of monozygotic twins.

The twin approach gave high estimates of the heritability and was fairly easy to obtain. But their heritability estimates were in effect the broad sense heritability and thus couldn't distinguish the additive genetic component and ultimately couldn't identify gene variants. Also, this approach couldn't adequately account for the effect of shared family environments which some argue might have inflated the heritability estimates obtained (Yang et al., 2010).

The heritability estimates obtained by the twin approach for all human traits, therefore, are higher bound estimates. These estimates have subsequently been shown to be higher than what could be explained by genetic markers like SNPs currently used in GWAS to estimate heritability (Clarke and Cooper, 2010; Maher, 2008; Manolio et al., 2009). SNP based estimates of heritability are mostly in the narrow-sense (Yang et al., 2010) depending on the family structure in the data and on whether the effects of siblings, for example, are accounted for (Xia et al., 2016).

The successful completion of the International HapMap Consortium project (Frazer et al., 2007) marked the end of twin studies and spurred the development of studies that targeted the population as a whole (Burton et al., 2007; McCarthy et al., 2008).

Initially, these studies utilized genotyping arrays developed from the more common SNPs among those discovered by the HapMap Consortium to genotype larger population samples. Recent studies utilize very dense genotyping arrays that incorporate rare SNPs. Those SNPs that appear on genotype arrays have been chosen

such that they are distributed as evenly as possible across the entire chromosomes of individuals (Burton et al., 2007).

These genotyped SNPs are rarely causal in terms of the traits being analysed but are located close to the causal variants that may directly affect these traits. Because variants that are close to each other on chromosome segments typically do segregate together more often than would happen due to chance (Reich et al., 2001), they show some correlation (Barrett, 2011). This non-random correlation is termed linkage disequilibrium (LD) and correlated SNPs are said to be in LD.

LD is the driving force behind genetic association studies using SNPs, since, for every variant that directly affects a trait, there might be genotyped SNPs nearby that are in LD with it, thus allowing genotyped SNPs to be associated with traits under study. The genetic association studies that exploit the LD between genotyped SNPs and ungenotyped possible causal variants are termed Genome-wide Association Studies (GWAS) because genotyped SNPs are selected to span the entire genome of individuals.

## 1.5 The genome-wide association study of complex traits

For a set of study individuals, the association between a trait of interest and variants at genomic loci is determined by regressing a vector of measured trait values on allele counts of SNPs genotyped for those individuals. The genotyped SNPs serve as proxies for all genomic loci. The genotyped SNPs are represented as a matrix of numerical codes taking the values of 0, 1 or 2 for each locus for each study individual. For each locus, there will be one of three possible genotypes derived from two

alleles. Each locus has a major and minor allele depending on their respective frequencies in the study population. The genotype codes used in GWAS are commonly derived as counts of the number of minor alleles at each locus. And once this is done, the minor alleles are termed as the effect alleles and their effects are estimated. This coding convention is commonly used because it makes evolutionary sense to assume that the allele with the smallest frequency arose recently by mutation and thus will have a differential effect on the trait. The major allele at the locus is then called the reference allele.

Three types of genetic models can be fitted for a GWA analysis. The most commonly used model is the additive genetic model which takes genotype coding 0, 1, and 2 for counts of the effect allele in the genotype. The second model is the dominant genetic model in which genotypes are coded as 0, 2, and 2 for the genotype with no effect allele, the genotype with one effect allele and the genotype with two effect alleles respectively. The third model is the recessive genetic model with genotype coding 0, 0, and 2 for the genotype with no effect allele, the genotype with one effect allele and the genotype with two effect alleles respectively. The dominant and recessive models are less used in GWA analysis.

The GWA analysis model fits the effect of each locus as fixed and regresses the vector of trait values on each locus one at a time starting from the first locus on chromosome one to the last locus on chromosome 22. Thus, a typical GWA analysis involves hundreds of thousands of regressions testing hundreds of thousands of genomic loci. The evidence of association between a trait and a genomic locus is

evaluated by p-values calculated for the null hypothesis of no association (Stephens and Balding, 2009).

The multiple testing of numerous loci presents a statistical problem of potentially having a lot of false positive associations for GWA analysis. This problem is dealt with by setting stringent thresholds for loci association p-values to pass before a locus is accepted as being genome-wide significantly (GWS) associated with a trait. This threshold has been calculated for European populations to be a p-value $< 1 \times 10^{-7}$ (Pe'er et al., 2008). However, a p-value $< 5 \times 10^{-8}$ is widely used now.

The first wave of GWA studies followed a case-control study design (Burton et al., 2007; McCarthy et al., 2008). These studies targeted binary traits which are defined by study participants answering yes or no to certain questions or having a status of disease or no disease. Then there was a shift towards population-based cohort studies that targeted continuous traits. These studies have been successful in improving our understanding of the genetic architecture of complex traits by making it possible to associate genomic regions to certain traits that affect populations (Visscher et al., 2012). And to some extent, the role the environment plays in these associations may also be shown for example in GWA studies that model gene-environment interactions (Thomas, 2010).

GWA studies, however, traditionally suffer potential drawbacks, some of which have been addressed. For instance, in GWA studies, it is commonplace for the study participants to be population stratified – individuals may not all come from the same ancestral or genetic population. Because of this, for some of the study

individuals, their allele frequencies for some of the genotyped SNPs will be different from the rest of the study population. Consequently, if the issue of population stratification is not adequately dealt with it may generate spurious associations in GWAS (Price et al., 2006). Because those stratified loci may be correlated with the trait, investigators may wrongly infer genetic association, whereas the association is, in fact, a result of individuals having different ancestral populations. Methods exist to detect and adjust for population stratification, so it is no longer a major issue.

But some inherent limitations still prevail in GWA studies. Among them is the issue of common SNPs on arrays only having modest effects sizes and explaining only a fraction of the genetic variation of the traits (Burton et al., 2007). There is also an issue of low power associated with GWA studies (Pe'er et al., 2006). The power of a GWA study is shown to be a function of the number of individuals sampled, the effect size of the genetic variants and the frequency of the alleles of the variants (Seng and Seng, 2008). The last two are mostly unknown until after the genetic variants have been uncovered and so the sample size remains the major controllable factor to improve power – the larger the sample size, the better the power (Burton et al., 2007; Sham and Purcell, 2014). Also, the extent of coverage of the genotyped SNPs affect power and so increasing the genome coverage of the SNPs (increasing the marker density) also improves power (Burton et al., 2007).

Although current GWA studies have a markedly improved power in terms of increased sample size and marker density (Gudbjartsson et al., 2015; Sudlow et al., 2015), they remain vulnerable to a range of errors and biases (McCarthy et al., 2008).

And some of these vulnerabilities contribute to the problem of the GWAS estimates of the genetic variation not fully accounting for heritability (Manolio et al., 2009).

## 1.6 Some of the heritability is "missing"

The science of explaining the genetic basis of trait variation, by the turn of the 21st century, had come to a point where we could for the first time, literally look into the human genome (the full DNA complement of an individual) to find the genetic component of common traits and diseases. By implementing analytical methods like the GWA analysis, researchers could map genomic loci that affect traits of interest and estimate the heritability of traits using loci that are genome-wide significant. But curiously, the heritability was not to be seen, it was apparently 'missing' (Maher, 2008). The heritability explained by genome-wide significant loci did not add up to that estimated from family studies. For instance, for a highly heritable trait like height which has heritability estimate of 80% from pedigree studies, the SNP heritability from genome-wide (GW) significant SNPs was less than 10% (Maher, 2008).

Finding the 'missing heritability' then became a priority in the field and stimulated a lot of discussion amongst population geneticists, human geneticists, animal geneticists and evolutionary biologists alike. Many investigators have since argued extensively as to where they think the problem of the missing heritability arises from and some propose how they think it could be resolved (Manolio et al., 2009).

The many possible ways that were proposed to potentially account for the missing heritability include, looking at copy-number variations (CNVs) (McCarroll, 2008); investigating the effects of non-additive effects such as epistatic interactions

amongst genes (Hemani et al., 2013); and also looking beyond sequence variation to look at the effect of epigenetic inheritance (Slatkin, 2009). Others reasoned that GWA studies which could capture rare gene variants with larger effects (Cirulli and Goldstein, 2010) and common gene variants with very small effects (Yang et al., 2010) could account for some of the missing heritability. Others suggested that adequately accounting for effects of close relatives by excluding one of each pair of relatives from analysis would provide more meaningful estimates of heritability explained by variants assayed (Yang et al., 2010), while others shared a different opinion and suggested models to adequately account for closely related individuals by including two different relationship matrices (one that accounts for all relationships and one that accounts for close relationships) in heritability calculations (Zaitlen et al., 2013). There is also the belief that perhaps we may be chasing 'phantom' heritabilities instead of 'missing' heritabilities because estimates of the total heritabilities could have been inflated by genetic interactions (Zuk et al., 2012).

The ensuing race to find the missing heritability led to the adoption of approaches that already existed and had been developed in animal breeding studies to try and predict the genetic values of individuals in breeding programs using genome-wide markers (Meuwissen et al., 2001; VanRaden, 2008). These techniques employ maximum likelihood approaches which fit Genetic Relationship Matrices (GRMs) in the estimation of the heritability from genome-wide SNPs. The technique was called Genomic-relatedness-matrix Restricted Maximum Likelihood (GREML) (Benjamin et al., 2012).

A study that used this GREML approach to study height found that all genome-wide SNPs, as opposed to GWS SNPs, explained more than 50% of the genetic variation in human height (Yang et al., 2010). This was a significant step in the search for the missing heritability. The success of this study led to the development of a tool, Genome-wide Complex Trait Analysis (GCTA) (Yang et al., 2011), that has since been used widely in complex traits heritability studies. Other tools that use the same approach have since been developed (Canela-Xandri et al., 2015). One such tool called Regional Heritability Advanced Complex Trait Analysis (REACTA) (Cebamanos et al., 2014) quantifies the local contribution of genomic regions to the total genetic variation of complex traits.

These GREML approaches generate lower-bound estimates of the narrow sense heritability as opposed to the estimates generated in twin studies (Benjamin et al., 2012). These GREML approaches have improved heritability estimates, but usually, do not model other structural variants like copy number variants and also cannot account for the non-additive components of the genetic variation. The latter may not be a big issue because the non-additive variation such as the dominance genetic variation has been shown to account for a very small proportion of the genetic variation of complex traits in humans (Zhu et al., 2015).

In the GREML approach, a Genomic Relationship Matrix (GRM) or kinship matrix is fitted to estimate the heritability. The relationship coefficients in these matrices are directly estimated from SNP genotype data. These relationship coefficients are normalised by per loci standard deviation which gives more weight to rare SNPs (Yang et al., 2010). A study that looked at these GRMs suggested

additional adjustment for local LD among SNPs (Speed et al., 2012). The argument was that some causal variants may be over-tagged (over-represented) by the genotyped variants due to their LD with them and thus may result in biased estimates of the additive genetic variation. But these LD adjustments have been argued not to work well in denser SNP arrays (Lee et al., 2013). There were also suggestions to employ different weights based on the minor allele frequency (MAF) of SNPs in the calculation of the GRM (Speed and Balding, 2015). These arguments led to an extension to the GREML model that clustered SNPs according to their MAF and regional LD (Yang et al., 2015).

Interestingly, amidst these arguments and counter-arguments, investigators have collectively not lost sight of the fact that full DNA resequencing of all samples remains the only way of knowing the full contribution of the genetic variation to the phenotypic variation (Bodmer and Bonilla, 2008; Cirulli and Goldstein, 2010; Manolio et al., 2009; Wray et al., 2013). Even with that, the contribution of some variants can be difficult to estimate individually because their individual effects are small and thus will need to be analysed collectively within a region.

### 1.6.1 Beyond the missing heritability and the rewards of trait prediction

Quantifying the genetic component of trait variation and the subsequent mapping of associated genetic loci is the primary aim of the genetic study of complex traits. However, this aim can be extended to trying to predict the unobserved phenotypes of a set of individuals based on the phenotypes observed for another set of individuals (Wray et al., 2013).

Trait prediction is useful in offering assistance in human genetic counselling and disease management. The heritability estimate of a trait conveys useful information like the extent to which the trait is under the control of genes. This information would influence actions aimed at reducing the prevalence of a disease. For instance, whether efforts should be made to identify genes involved and make further efforts to know the mechanisms by which these genes influence disease susceptibility or perhaps just look at environmental contributions for answers. Mapping disease risk gene loci can be used in diagnosis to predict the chances of an individual carrying risk alleles developing the disease in the future.

In agriculture, trait prediction is useful for predicting progeny performance in plant and animal breeding. The heritability estimate of traits helps breeders to make informed guesses about the outcome of breeding experiments. For instance, what to expect if only individuals with a trait value greater than a certain threshold are allowed to breed. This provides significant economic gains.

## 1.7 Other genetic markers (haplotypes)

In diploid organisms, there are two alternative SNP alleles at any given genomic locus. These two alternative alleles form the genotype at that locus. A set of linked SNP alleles on the same chromosome is called a haplotype.

Haplotypes are formed within haplotype blocks on a chromosome. The haplotype blocks arise due to recombination on chromosomes and thus these blocks are bounded by recombination hotspots (Daly et al., 2001). Within a haplotype block,

there is little, or no recombination occurring (Frazer et al., 2007). Therefore, the SNPs within a haplotype block tend to be inherited together.

The SNP alleles within a haplotype block combine differently for individuals in the population to generate several haplotype patterns. For a haplotype block containing two SNPs, then there can be four possible combinations of the alleles to generate four haplotypes. For a diploid individual, there can be nine possible haplotype pairs at this two-SNP haplotype block. Any pair of haplotypes is called a diplotype.

These diplotypes can be used analogously to genotypes to map genomic loci associated with traits of interest. In a GREML setting, diplotypes can be used to calculate relationship-values between individuals which can be used to estimate the genetic variance.

## 1.8 Aims of this study

One thing which is becoming increasingly clear from all the GWA studies and the missing heritability debates is that the way genetic associations are assessed by investigators is critical and has a huge role to play in the estimation of genetic variation. Naive estimations of these genetic associations may lead to erroneous results, which in effect would impact on our ability to unravel the genetic underpinning of complex traits. This study, therefore, aims to critically explore the statistical models used in determining genetic association and more importantly, highlight how violations of the model assumptions impact model estimates. The study also aims to develop analytical methods that incorporate other genetic variants

such as haplotypes and test these models on simulated datasets and human population data collected by ourselves and our collaborators.

This project uses large human datasets containing dense SNP genotypes and phased data along with phenotypic records to investigate how genetic information should be used to quantify the genetic component of variation and to determine causative genetic variants and how they contribute to the phenotype variation.

The thesis contains five data chapters that seek to address the aims of this study. Chapter two introduces the basic model used in heritability estimation and association analysis using packages in R and specialized genetic software such as PLINK (Purcell et al., 2007), GCTA (Yang et al., 2011), REACTA (Cebamanos et al., 2014) and DISSECT (Canela-Xandri et al., 2015). This chapter explores these models using urine phenotypes collected from individuals in the Generation Scotland: Scottish Family Health Study (GS: SFHS) (Smith et al., 2006).

Chapters three and four explore the merits of utilising Bayesian analytical approaches in a genome-wide association setting. Chapter three uses simulated complex phenotypes to explore the Bayesian model and makes a comparison to the Genomic Best Linear Unbiased Predictor (GBLUP) method. Chapter four tests the Bayesian model on urine phenotypes from GS: SFHS individuals.

Chapters five and six implement a regional GREML model that analyses the genome in natural blocks delimited by recombination hotspots. Chapter five tests the regional GREML model in a simulation study and chapter six applies the model to human complex trait data from GS: SFHS and UK Biobank (Sudlow et al., 2015).

The concluding chapter of the thesis discusses the wide-reaching implications of all the results of the data chapters to the genetic study of human complex traits and discusses directions and strategies for future studies in the field.

# Chapter 2

## 2 Investigating the genetic control of urinary traits in individuals in the Scottish population using a linear mixed model

### 2.1 Introduction

Kidney failure is a global public health challenge with several million people requiring treatment (Xue et al., 2001). This presents a distressing economic burden on health care systems and thus raises the need for research aimed at reducing the incidence of and severity of renal failures among people. One of such research approaches is the Genome-wide Association (GWA) study, which for the past decade has been incredibly successful in stacking up evidence to suggest that genetic predisposition plays a crucial role in the aetiology of complex human diseases by detecting trait-locus associations.

The GWAS approach, so far, has been a robust tool for the study of complex traits related to kidney function and disease – with a myriad of GWA studies attesting to this fact (Piras et al., 2017). For example, Chambers et al. (2010) identified five loci associated with serum creatinine, a marker of kidney function. Also, genome-wide

scans in over 90,000 Caucasian individuals by Köttgen et al. (2010) identified 20 novel genome-wide significant loci for reduced renal function and CKD. Collectively, these and many more GWAS findings (Hishida et al., 2018; Wuttke et al., 2016) have shed some light on kidney function and the aetiology of kidney disease.

Measures of urine electrolytes provide useful information on kidney function in people and are used routinely by clinicians as biomarkers of kidney damage (Wallace et al., 2008). For instance measures of creatinine are used to calculate functional excretion of $Na^+$ and renal failure index (Reddi, 2014). Such measurements are intermediate phenotypes that are associated with the disease but not by themselves symptoms of the disease and are heritable (Cirulli and Goldstein, 2010).

The use of such intermediate phenotypes (endophenotypes) has been proposed as a more amenable alternative to the actual disease itself in assaying for genetic risk variants (Hall and Smoller, 2010). The reason for this is that these intermediate phenotypes may be closer to the underlying genetic liability than the actual disease (Almasy and Blangero, 2001; Ghosh and Collins, 1996). Another reason is that endophenotypes may be clinically and ethnically more homogeneous than disease endpoints (Ghosh and Collins, 1996) and thus help to deal with the problem of heterogeneity in disease status ascertainment. Wrongly defining cases and controls for a disease can result in spurious genetic associations which can make it impossible to replicate the results in other cohorts. GWA scans to identify significant associations in kidney disease endophenotypes like urine measures of potassium, sodium, calcium etc. and subsequent estimates of heritability, potentially, will at a first instance improve our understanding of the genetic underpinning of renal

pathogenesis which ultimately will translate into providing new insights into kidney disease biology. There has been a GWA study of urine sodium, potassium and creatinine (Wallace et al., 2008), however, no genome-wide significant association was identified.

On this premise, this study set out to investigate the genetic control of urine electrolyte measures of 3,000 individuals from the Generation Scotland: Scottish Family Health Study (GS: SFHS) (Smith et al., 2006). The GS: SFHS involves about 24,000 individuals recruited from across Scotland and have measurements for a wide variety of health-related phenotypes taken on them. Among these phenotypes are measurements for 8 urine electrolytes measured for about 3,000 individuals. The aim was to perform a mixed model GWAS on the individuals for all the 8 urine traits in a bid to uncover genetic risk loci underlying these kidney disease endophenotypes and then estimate genome-wide and regional heritabilities to understand how the underlying genetic variation could influence the inheritance of these traits in a way that will inform future research directions on renal function and failure.

## 2.2 Methods

### 2.2.1 The Generation Scotland kidney phenotypes dataset

The data comprised of 2,934 individuals from Glasgow (2,363) and Aberdeen (571) which formed part of a larger dataset of 10,000 out of the 23,960 individuals that were initially genotyped in the Generation Scotland: Scottish Family Health Study (GS: SFHS) (Smith et al., 2012). DNA from the individuals had been analysed using the Illumina HumanOmniExpressExome8v1-2_A chip (~700K genome-wide SNP chip). These individuals had phenotype measures for the urine concentrations of,

Sodium, Potassium, Chlorides, Calcium, Glucose, Phosphorus, Magnesium, urine osmolarity and Creatinine. The phenotype data included the covariates sex, age and Clinic location (which was a vector of numbers representing recruitment centres in Aberdeen and Glasgow).

Standard QC filters were applied to the genotype data in GenABEL (Aulchenko et al., 2007) and also in PLINK version 1.9 for the PLINK version of the data. The QC excluded SNPs and individuals with a call rate of less than 98%, SNPs with MAF<0.01 and SNPs that were out of Hardy-Weinberg equilibrium (p-value 0.000001). All the individuals and 562,273 SNPs successfully passed all QC filters and were used in analysis downstream.

### 2.2.2 Principal components analysis to check for population substructure and ancestry

The resulting data after applying the QC filters were used to investigate the level of population stratification and ancestry of the individuals. The analysis to check for population substructure was done in R using an approach which involved the calculation of a covariance matrix $\boldsymbol{G}$ of all individuals in the sample using all the genome-wide markers. The resulting matrix is positive semi-definite, and thus decomposes into a diagonal matrix $\boldsymbol{D}$ which has non-negative elements and an orthogonal matrix $\boldsymbol{X}$ (Freedman, 2009) such that

$$G = XDX^{I} \tag{2.1}$$

The columns of $\boldsymbol{X}$ represent the eigenvectors of $\boldsymbol{G}$ and the diagonal elements of the matrix $\boldsymbol{D}$ are the eigenvalues of $\boldsymbol{G}$. The principal components analysis (PCA) is performed such that the first two principal components (PCs), after a

multidimensional scaling of $\boldsymbol{G}$, are computed and returned. The principal components are the columns of $\boldsymbol{X}$ or the eigenvectors of $\boldsymbol{G}$. The first component is the column that generates the greatest variance among the study subjects after a linear arrangement of the columns of $\boldsymbol{X}$ i.e. the eigenvector with the largest eigenvalue. The variances of the components shrink as one moves away from the first PC. A plot was generated from the first two PCs.

A second plot was generated to investigate the ancestry of GS: SFHS individuals using analyses in PLINK (version 1.9) that merged the individuals of the GS: SFHS data with the individuals from 14 ancestral populations (Table 2.1) in the 1000 Genomes (1kG) project (Consortium, 2012).

Table 2.1. The 14 different populations from the 1000 Genomes Project that were used as reference ancestral populations in this study

| Population Code | Population Description | Super Population Code |
|---|---|---|
| **CHB** | Han Chinese in Beijing, China | EAS |
| **JPT** | Japanese in Tokyo, Japan | EAS |
| **CHS** | Southern Han Chinese | EAS |
| **CEU** | Utah Residents (CEPH) with Northern and Western European Ancestry | EUR |
| **TSI** | Toscani in Italia | EUR |
| **FIN** | Finnish in Finland | EUR |
| **GBR** | British in England and Scotland | EUR |
| **IBS** | Iberian Population in Spain | EUR |
| **YRI** | Yoruba in Ibadan, Nigeria | AFR |
| **LWK** | Luhya in Webuye, Kenya | AFR |
| **ASW** | Americans of African Ancestry in SW USA | AFR |
| **MXL** | Mexican Ancestry from Los Angeles USA | AMR |
| **PUR** | Puerto Ricans from Puerto Rico | AMR |
| **CLM** | Colombians from Medellin, Colombia | AMR |

The PLINK analyses involved pruning the genotype data of the merged dataset based on LD. The LD-thinned data was used to compute a covariance matrix which

was subsequently transformed into a distance matrix in PLINK. The first two PCs were plotted in R to show how the GS: SFHS individuals clustered with respect to the individuals from different ancestral populations.

### 2.2.3 Phenotype data transformation and regression to generate residuals

The residuals for the kidney function traits were obtained after regressing the trait values on the covariates; age, sex, clinic location and creatinine. For a vector of phenotypes with length $n$, the linear regression model for fitting the effects of measured covariates is given as

$$y = 1\mu + X\beta + e \qquad (2.2)$$

where $\boldsymbol{y}$ is an $n \times 1$ vector of measured traits. $\mathbf{1}$ is a vector of 1s and $\mu$ is the trait mean. $\boldsymbol{X}$ is an $n \times p$ design matrix of the fitted covariates. Creatinine was fitted in the regression model as a quadratic function. $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects and $\boldsymbol{e}$ is the residual error which is random. To estimate $\boldsymbol{\beta}$ (the effects of the fitted covariates on the measured trait values) from the data, the linear regression model makes some assumptions that link the model to the data. First the model assumes that the expectation of $\boldsymbol{y}$, ($E(\boldsymbol{y})$ is effectively described by a simple observed function of the fitted variables on the right hand of equation (2.2) (linear function for age, sex and clinic location, and a quadratic function for creatinine) (i.e. the values of the vector $\boldsymbol{y}$ are observed values of $1\mu + X\beta + e$) (Freedman, 2009). Second is the assumption of normality in the distribution of the residuals and third is the assumption of no relationship between the expectation and the measure of the dispersion of data points around this expectation (i.e. constant variance with changing mean).

If these assumptions are not met, then transforming the values of $\boldsymbol{y}$ may provide some improvement to the solution (Box and Cox, 1964). The model diagnostic plots provide a means to visually assess these assumptions to know if the model fits the data well enough. These are plots of Residuals versus fitted values, Normal Q-Q plot, Scale-Location plot and the residuals versus leverage plots.

In this study, kidney function traits in which the model was not well behaved were transformed using Box-Cox transformations (Box and Cox, 1964) in R. The Box-Cox transformation of the $ith$ value of $\boldsymbol{y}$ is defined as

$$y_i^{\lambda} = \frac{y_i^{\lambda} - 1}{\lambda} \; where \; \lambda \neq 0 \qquad (2.3)$$

where $y_i$ is the $ith$ value of the vector of observations and $\lambda$ is the transformation parameter. For $\lambda = 0$, the natural log of $\boldsymbol{y}$ is taken in place of equation (2.3) i.e. $y_i^{\lambda} = loge(y_i) \; where \; \lambda = 0$ (Osborne, 2010).

Urine concentrations of Chlorides, Potassium, Sodium and Osmolarity were transformed using $\lambda = 0.43, 0.3, 0.383 \; and \; 0.64$ respectively. Urine Calcium had $\lambda = 0$ and thus was transformed using the natural log. Urine Phosphorus and Magnesium were transformed using the log10. Urine Glucose was transformed using quantile normalisation, after all the transformation methods described above failed to normalise the trait values. The residuals of both the transformed and untransformed phenotype data were used in the downstream GWA and heritability analysis.

### 2.2.4 GREML analysis

The heritability is the genetic component of the phenotypic variance. For a vector $\boldsymbol{y}$ of $n$ trait values, the mixed effects model for fitting the polygenic effect of genome-wide SNPs is given as

$$y = X\beta + Zu + e \tag{2.4}$$

where $\boldsymbol{y}$ is an $n \times 1$ vector of phenotypes, $\boldsymbol{X}$ is an $n \times k$ design matrix of $k + 1$ fixed effects, and $\boldsymbol{\beta}$ a $k \times 1$ vector of fixed effects, $\boldsymbol{Z}$ is an $n \times m$ design matrix for polygenic effects and $\boldsymbol{u}$ is an $m \times 1$ vector of random polygenic effects of $m$ genome-wide SNPs assumed to be multivariate normal, $MVN\left(0, \sigma_g^2 \boldsymbol{G}\right)$. And $\boldsymbol{e}$ is an $n \times 1$ vector of residual effects assumed to be multivariate normal, $MVN(0, \sigma_e^2 \boldsymbol{I})$.

Under this model, the vector of traits is assumed to be normally distributed, $N(\boldsymbol{X\beta}, \boldsymbol{V})$. And the variance, $V$, for a trait is defined as

$$V = G\sigma_g^2 + I\sigma_e^2 \tag{2.5}$$

where $\boldsymbol{G}$ and $\boldsymbol{I}$ are the incidence matrix for the total additive genetic variance $\sigma_g^2$ and residual environmental variance $\sigma_e^2$ respectively. Both $\sigma_g^2$ and $\sigma_e^2$ are estimated with GCTA (Yang et al., 2011) using Restricted Maximum Likelihood (REML) where the Genetic Relationship Matrix, $\boldsymbol{G}$, is fitted.

The matrix $\boldsymbol{G}$ is expressed as

$$G = \frac{ZZ'}{m} \tag{2.6}$$

where $\boldsymbol{G}$ is the genetic relationship matrix (GRM) between pairs of individuals which is estimated as the proportion of the genome-wide autosomal SNPs shared Identical by State (IBS) across all $m$ markers weighted by their allele frequency (Yang et al., 2010); $\boldsymbol{Z}$ is a matrix that indicates which marker alleles each sampled individual carries and has been centred and standardized at each locus (VanRaden, 2008); $m$ is the total number of markers used. The genomic relationship coefficient for two individuals $i$ and $k$ is therefore estimated as follows

$$G_{ik} = \frac{1}{m} \times \sum_{j=1}^{m} \frac{(x_{ij} - 2p_j)(x_{kj} - 2p_j)}{2p_j(1 - p_j)} \quad (2.7)$$

where $x_{ij}$ is the genotype code at locus $j$ for individual $i$ and takes the values 0, 1 and 2 for $AA, Aa$, and $aa$ genotypes respectively, $p_j$ is the frequency of allele $a$ at locus $j$.

The narrow sense heritability, $h^2$ is then calculated by

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2} \quad (2.8)$$

This gives an estimate of the heritability that is explained by genome-wide SNP markers (Yang et al., 2011).

The GREML analysis of the GS dataset was performed using the transformed data. Two sets of analysis were performed for each category of the dataset. The first was an analysis that excluded one of each pair of relatives (individuals with estimated kinship > 0.025) and computed a global estimate of heritability. The second analysis fitted separate relationship matrices for all the 22 different chromosomes (autosomes) simultaneously in the GREML analysis to try and estimate the local

contribution of each chromosome to the total heritability. The model for the chromosome GREML is given as

$$y = X\beta + \sum_{i=1}^{22} Z_i u_i + e \qquad (2.9)$$

### 2.2.5 Mixed effects linear model to test for association

The association between SNPs and a trait was tested using a mixed effects linear regression model,

$$y_r = 1\mu + Xg + Zu + e \qquad (2.10)$$

where $\boldsymbol{y_r}$ is a vector of the residuals for the trait after adjusting for the covariates, $\mu$ is the mean effect which is fixed; $\boldsymbol{g}$ is a vector of allelic substitution effect, which is fixed with a design matrix $\boldsymbol{X}$ (the matrix $\boldsymbol{X}$ takes the values 0, 1, or 2 for the $AA, Aa$ and $aa$ genotypes respectively); $\boldsymbol{u}$ is a vector of random polygenic effects with a design matrix $\boldsymbol{Z}$; $\boldsymbol{u}$ is assumed to be multivariate normal, $\boldsymbol{u} \sim MVN\left(0, \boldsymbol{G}\sigma_g^2\right)$; $\boldsymbol{e}$ is the residual environmental effect, also assumed to multivariate normal, $\boldsymbol{e} \sim MVN(0, \boldsymbol{I}\sigma_e^2)$.

$\boldsymbol{G}$ was calculated in the GenABEL package in R (Aulchenko et al., 2007) using the "***ibs"*** function. The $\boldsymbol{G}$ matrix was subsequently used in the "***polygenic"*** function (Thompson and Shaw, 1990) to estimate the heritability for the kidney disease traits already adjusted for the fixed effects (sex, age, clinic location and creatinine). The results from the polygenic function were saved as R objects which were later employed in the "***mmscore"*** function (Chen and Abecasis, 2007) in GenABEL to estimate the association between SNPs and traits by the mixed effects linear regression model described in equation (2.10).

### 2.2.6 Zoom in around top GWAS hits

A further exploratory analysis that focused on the top GWAS hit SNPs for the kidney function traits was performed by selecting SNPs within 500kb on both sides of those SNPs and fine mapping using LocusZoom (Pruim et al., 2010). LocusZoom is an online plotting tool that provides detailed regional plots around GWAS hits. The plots list genes in the region, local LD between SNPs and recombination rates.

### 2.2.7 Regional GREML analysis

Consider a vector $\boldsymbol{y}$ of phenotype values with length $n$, the mixed effects model for fitting the effects of $r$ genomic regions and $m$ background polygenic markers is given as

$$y = X\beta + W_i u_i + Z u_p + e \qquad (2.11)$$

where $\boldsymbol{y}$ is an $n \times 1$ vector of phenotypes, $\boldsymbol{X}$ is an $n \times k$ design matrix of $k + 1$ fixed effects, and $\boldsymbol{\beta}$ a $k \times 1$ vector of fixed effects, $\boldsymbol{W_i}$ is an $n \times s$ design matrix relating phenotype measures to $s$ SNPs in region $i$ and $\boldsymbol{u_i}$ is an $s \times 1$ vector of random genetic effects due to region $i$ assumed to be multivariate normal, $MVN\left(0, \sigma^2_{u_i}\boldsymbol{G_{u_i}}\right)$. $\boldsymbol{Z}$ is an $n \times m$ design matrix for $m$ background polygenic effects of SNPs outside the region and $\boldsymbol{u_p}$ is an $m \times 1$ vector of random polygenic effect of these SNPs excluded from the regions assumed to be multivariate normal, $MVN\left(0, \sigma^2_{u_p}\boldsymbol{G_{u_p}}\right)$. And $\boldsymbol{e}$ is an $n \times 1$ vector of residual effects assumed to be multivariate normal, $MVN(0, \sigma^2_e \boldsymbol{I})$.

Under the model, the vector of phenotypes $\boldsymbol{y}$ is assumed to be normally distributed, $N(\boldsymbol{X\beta}, \boldsymbol{V})$ where the variance is

$$V = \sigma^2_{u_i} G_{u_i} + \sigma^2_{u_p} G_{u_p} + \sigma^2_e I \qquad (2.12)$$

The regional GREML analysis was performed using REACTA (Cebamanos et al., 2014). This program performs a regional GREML analysis across the whole set of autosomal SNPs by breaking up the genome into smaller windows of N SNPs, which overlap by X SNPs. For this study, the analysis was carried out using a window size of 100 SNPs, and consecutive windows overlapped by 50 SNPs. For each SNP window, the local GRM for that window was analysed simultaneously with a whole genome GRM that excludes the SNPs in the region.

## 2.3 Results

### 2.3.1 PCA places GS: SFHS individuals in European ancestry

The generation of spurious genomic associations in disease studies arising from population stratification is a well-known fact in GWAS literature (Aulchenko, 2011; Freedman et al., 2004). Techniques exist to deal with this problem with genomic control (Aulchenko, 2011) being the most widely employed. But these have their limitations (Price et al., 2006). In this study, population substructure was checked for by computing the first two principal components from a covariance matrix of study participants calculated using all the genome-wide markers.

Figure 2.1a shows the results of this analysis. The plot shows most individuals in the GS: SFHS data clustering together on the first PC axis with very few individuals breaking away from the cluster. The second PC, however, shows a big gap between seven individuals (shown in red) and the rest of the population. Further investigation of these individuals revealed that they belong to the same family that spans three generations from grandparents to grandchild. There is not much substructure in the GS: SFHS population apart from the one produced by this family. Keeping them in the

data for GWA analysis shouldn't cause any problems because the mixed model used fitted a GRM to account for background genetics and this should adequately deal with any problems that keeping these individuals might cause.

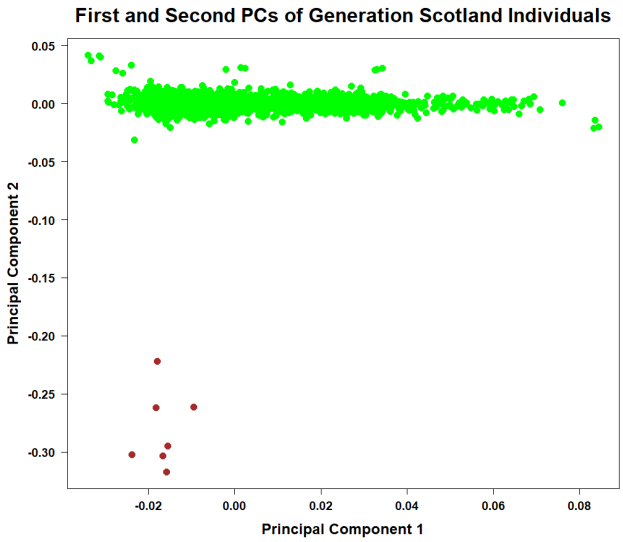
The analysis to check for ancestry firmly placed the GS: SFHS individuals amongst individuals of European ancestry (Figure 2.1b). Both the first and second principal components clearly separated the GS: SFHS individuals from the African and Asian populations, confirming the European ancestry of the GS: SFHS individuals. The populations from the Americas (ASW, MXL, PUR, CLM) were spread in between the European (CEU, TSI, FIN, GBR, IBS), Asian (CHB, JPT, CHS) and African (YRI, LWK, ASW) populations.

### 2.3.2 Data transformations improve model fit

By transforming data, we do two things. Firstly, we change the nature of the relationship between the data variables. This affects our interpretation of results because it may change the effect sizes in whichever direction, up or down. Second is that we alter the mean-variance relationship which again makes results interpretation difficult. Albeit transformations attempt to minimize violation of model assumptions which can improve the results from the analyses by minimizing our chances of committing either Type I or II errors (Osborne, 2010).

The results for the non-linear transformations of the data using λ values from the Box-Cox analyses are shown in Figure 2.2.

**a.**




**b.**

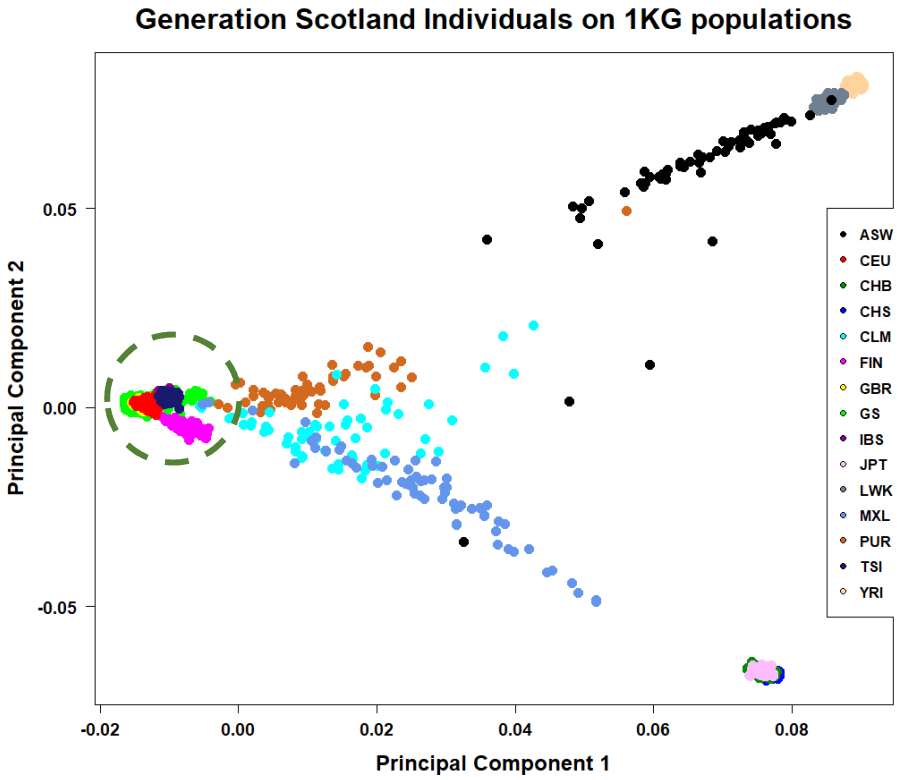



Figure 2.1. Plot of the first two principal components of covariance matrix of study individuals. a. plot for the GS: SFHS data only. The outlier points shown in red are individuals from the same family spanning 3 generations. The plot shows there is not much genetic substructure in the GS data; b is a plot of the GS: SFHS individuals with the individuals from 14 ancestral populations from the 1000 Genomes project. The GS: SFHS individuals clustered together with individuals of European ancestry (dashed circle). They were clearly separated from the African Asian ancestral populations on both PC axes with the populations from the Americas spread in between.

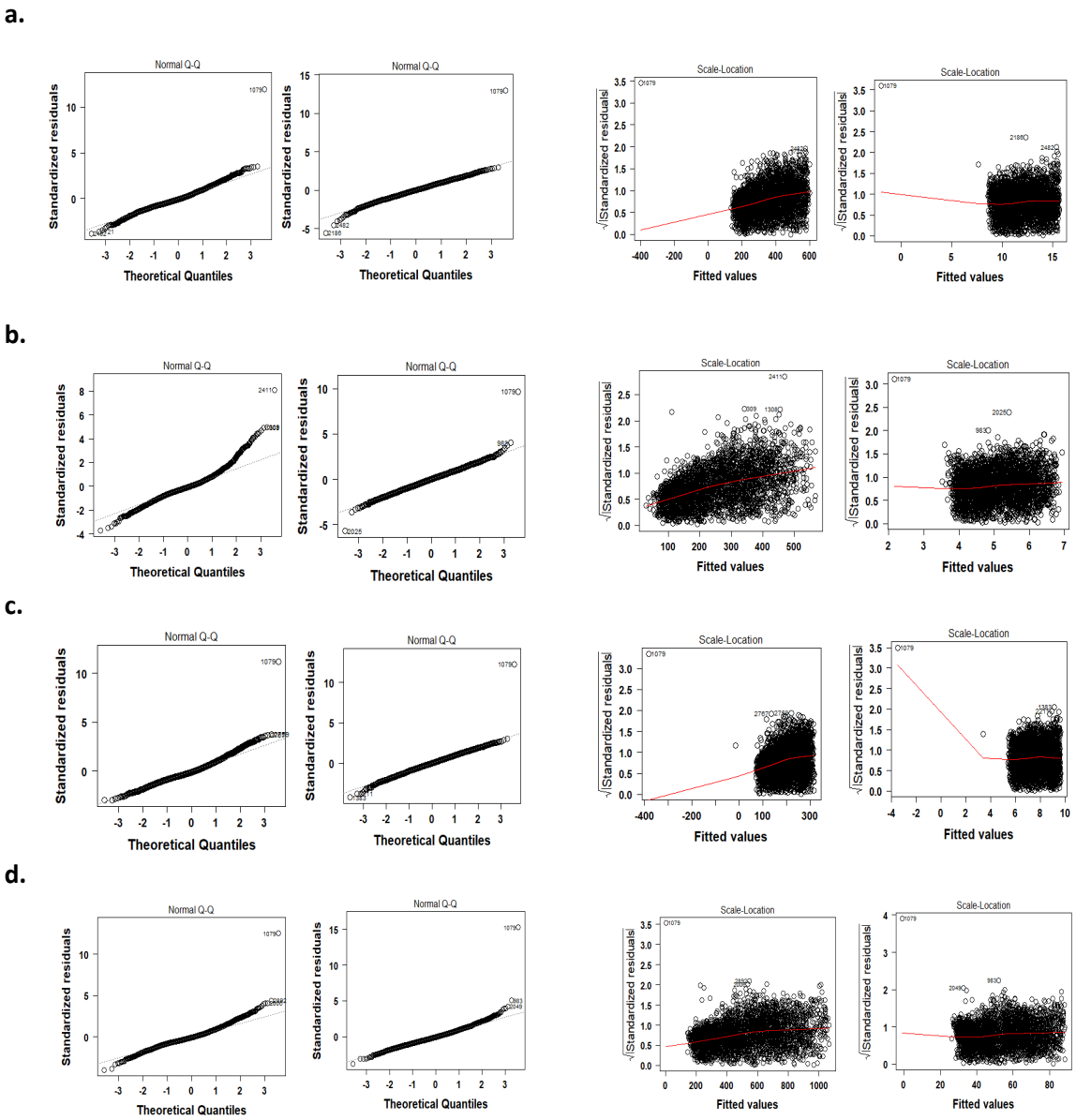



Figure 2.2. Model diagnostics plots before and after Box-Cox transformations. From a to d is urine chlorides, potassium, sodium and osmolarity respectively. Each phenotype has got 4 model diagnostic plots. The first two tests for the assumption of normality in the distribution of residuals before and after Box-Cox transformations. The other two plots show the mean – variance relationship before and after transformation. The Box-Cox transformation significantly improves the model fit and minimizes violations of model assumptions

This is a plot of model diagnostics before and after Box-Cox transformations for urine chlorides, potassium, sodium and osmolarity. The results in Figure 2.2 show some improvement in two of the model assumptions. First is the assumption of

normality in the distribution of residuals which is shown in the figure by the Q-Q plots. The Q-Q plots in the transformed data show the plotted points falling closely on the diagonal line. The improvement is quite obvious in urine potassium (Figure 2.2b) where the Q-Q plot of the residuals in the untransformed data shows that the residuals are skewed in such that the larger values are more extreme than would be expected under a normality, thus skewing the data to the left. This is corrected by the Box-Cox transformation (Figure 2.2b).

The second improvement is the assumption of constant variance with increasing mean. The plots for Box-Cox transformed data shows parallel dispersal of points as we would expect for the assumption of no relationship between the variance and the mean. The plot for urine sodium (Figure 2.2c) shows a distortion of the dispersion by a single outlier point.

The log10 transformed phenotypes also showed some improvement over the untransformed data with respect to the assumptions of normality in distribution of residuals and constant variance with increasing mean, Figure 2.3. None of the transformation tested improved the distribution properties for glucose and therefore quantile normalisation was applied in further analyses.

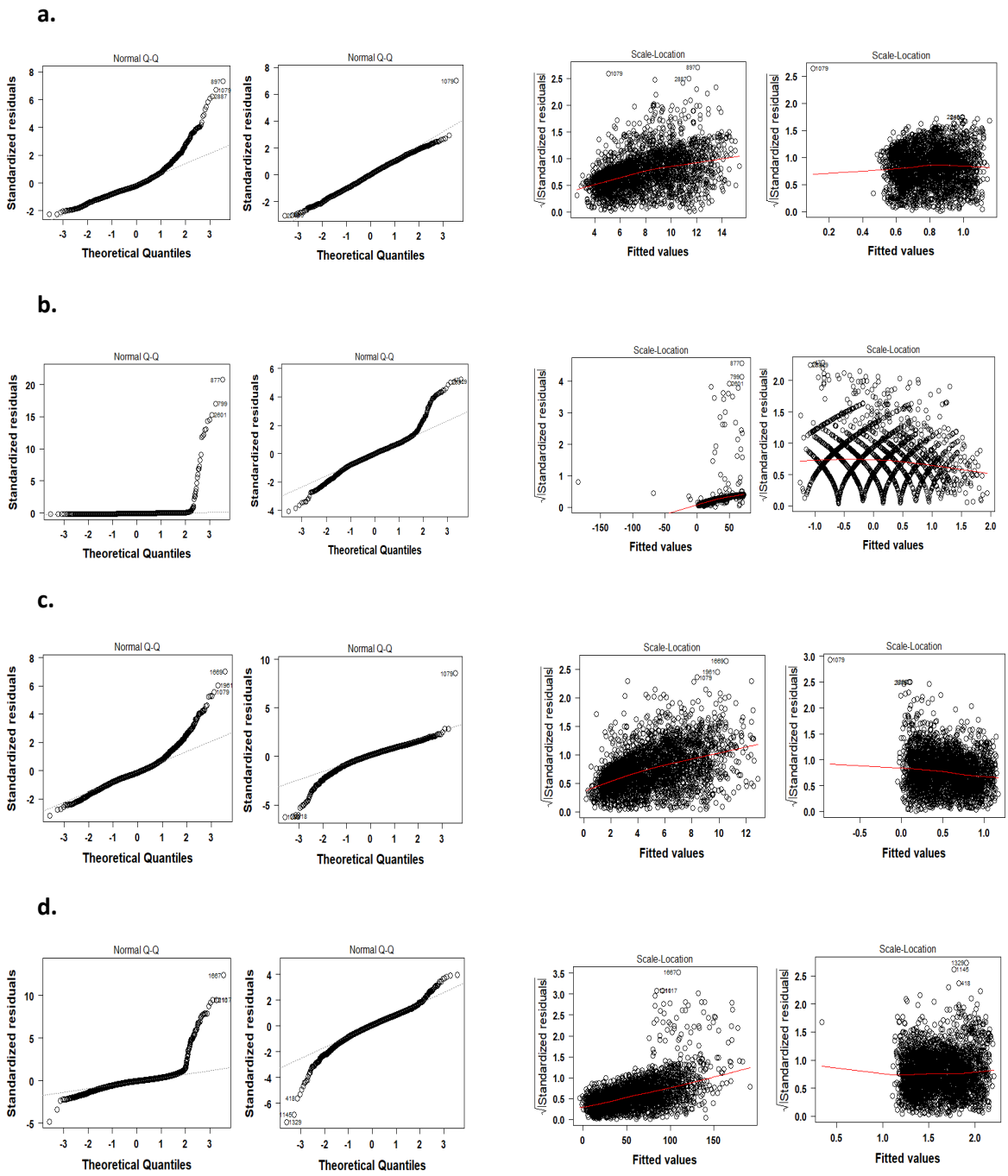



Figure 2.3. Model diagnostics plots before and after log10 transformations. From a to d is urine calcium, glucose, magnesium and phosphorus respectively. The log10 transformation improves the model fit especially for calcium but doesn't do as well as the Box-Cox transformations in magnesium and phosphorus. The quantile normalisation of urine glucose offered some improvement on the residual distribution of the trait.

### 2.3.3 GREML analysis to estimate heritability

The results for the GCTA analysis are shown in Table 2.2 and plots for the chromosomal heritability of the traits are shown in Figure 2.4 and Figure 2.5. The heritability estimate of a trait gives an idea of whether the trait is significantly affected by genes or not. Higher estimates are an indication of significant contribution by genes to trait variation. The whole genome estimates of heritability were obtained for the kidney phenotypes by analysing all common SNPs in the genome at once using restricted maximum likelihood analysis. This gave rather small estimates of heritability for most of the traits.

Table 2.2. The heritability estimates of the two categories of GS: SFHS data. The trait name, the heritability estimates for traits, and the sum of chromosomal heritability estimates of traits. The standards errors are in parentheses. NC is for likelihood not converged.

| Traits | Heritability Estimates | |
|---|---|---|
| | Whole Genome $h^2$% (SE) | Sum of Chromosomal $h^2$% (SE) |
| Calcium | 5.76 (6.2) | 20.74 (7.00) |
| Chlorides | 1.25 (5.51) | 9.70 (6.11) |
| Glucose | 0.00 | NC |
| Potassium | 0.00 | NC |
| Magnesium | 10.50 (5.73) | 16.29 (6.00) |
| Sodium | 10.52 (5.73) | 14.92 (5.87) |
| Osmolarity | 4.22 (5.52) | 13.07 (6.02) |
| Phosphorus | 0.00 | NC |

When the SNPs were binned by chromosomes and local (chromosomal) estimates of heritability were calculated and summed, all the traits had an increase in the heritability estimates (Table 2.2). The likelihood failed to converge for the chromosomal estimates for some traits after 1000 iterations and they have NC (Not Converged) entered in Table 2.2. The standard errors of the estimates are large, but this is likely to improve with the increase in sample size. The regional heritability

analysis further reduced the local bin sizes from whole chromosomes to 100 SNP bins overlapping by 50 SNPs. The results are shown in Figure 2.4b and Figure 2.5b. The results for regional heritability analysis for most of the traits showed some consistency with the chromosomal heritability results.

### 2.3.4 Mixed model GWA analysis identify genome-wide significant loci

The most important result from this analysis is that there are significant GWAS hits in most of the kidney traits in areas of the genome that have genes within 1MB (Table 2.3 and Table 2.4). The majority of these hits, however, are significant at the suggestive level and do not reach genome-wide significance. But concerns over the overly stringent threshold of genome-wide significance (p-value $< 5 \times 10^{-8}$) have been raised previously (Sham and Purcell, 2014). Although the reason for this threshold is mainly statistical and may lack any biological significance, its usage presents the possibility of true associations of biological relevance being missed by GWA studies. Looking at SNPs that are significant at the suggestive level (p-value $< 1 \times 10^{-5}$) is now common practice to mitigate this concern. But beyond this point, there is not much motivation to look any further.

Another important observation from these analyses is that the data transformations greatly reduced the SNP effect sizes (Table 2.3 and Table 2.4). This doesn't come as surprise because transformations obviously change the scale of the data and in effect should change the scale of the effect sizes. Data transformation, however, improved the evidence of association by giving low association p-values. The GWA results for the traits are shown in the Manhattan plot in Figure 2.4c and Figure 2.5c.

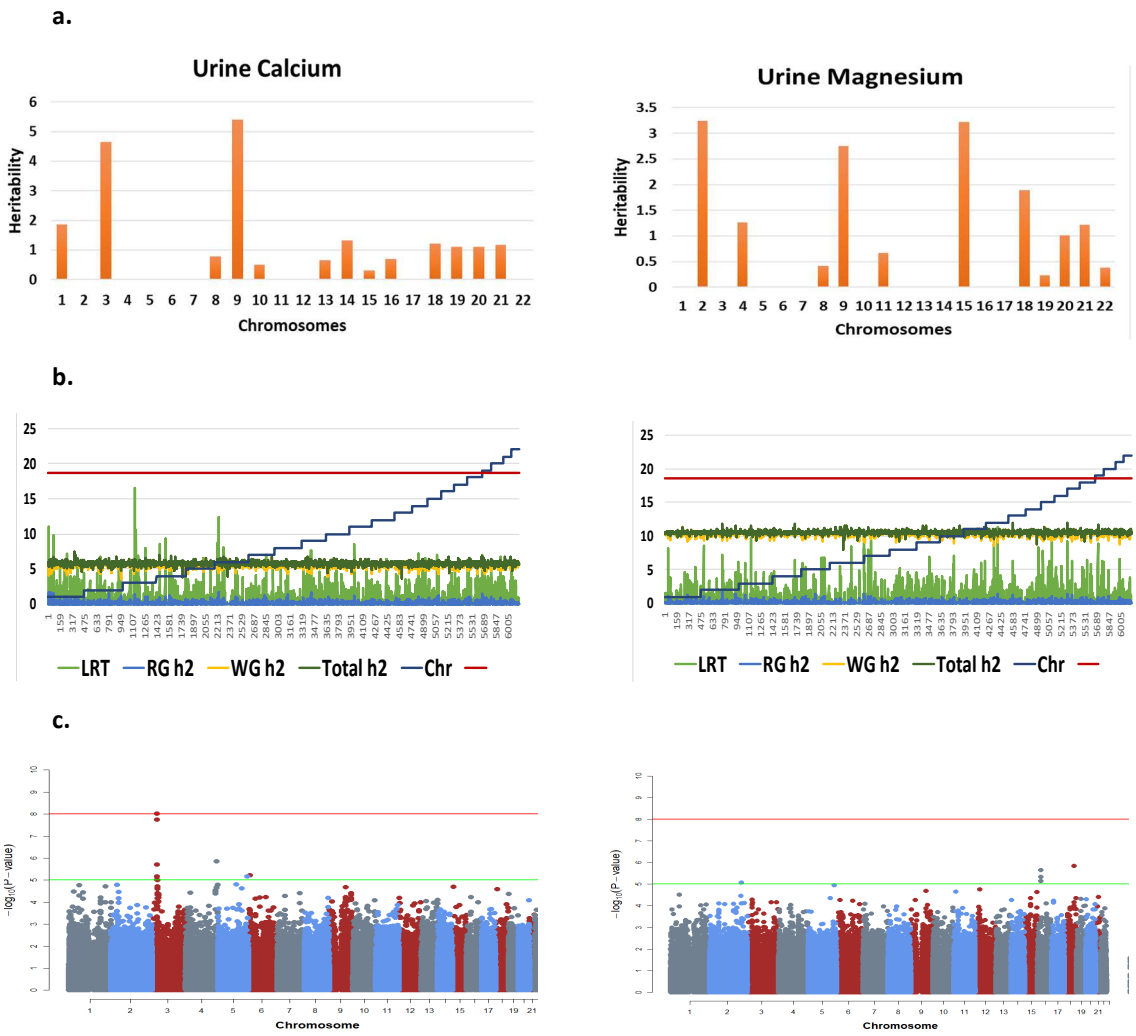



Figure 2.4. Whole genome and regional analysis of urine calcium and magnesium. **a** plots of chromosomal heritability, with the heritability of chromosomes expressed as percentage**. b** plots of regional GREML analysis. The Likelihood ratio test statistic (LRT), regional heritability (RG h2), heritability of rest of genome excluding region (WG h2) and total heritability (Total h2) are plotted. The red horizontal line is Bonferroni-corrected threshold for genome-wide significance (LRT=18.6, p-value < 1.6278e-05). Chromosomes are indicated by blue steps plot. **c** Manhattan plots showing genome-wide associations. The points are plots of -log10 of association p-values of genome-wide SNPs. Red line is genome-wide significance (GWS) threshold (p-value < 1e-08) and green line is threshold for significance at the suggestive level (p-value < 1e-05). For calcium the Manhattan plot shows a GWS peak on chromosome 3 where one SNP reaches genome-wide significance. The regional GREML plot shows the most significant region on chromosome 3 (LRT = 16.553, p-value =2.37e-05). The chromosomal heritability plot points to chromosome 3 as explaining the second largest proportion of the total heritability of calcium.

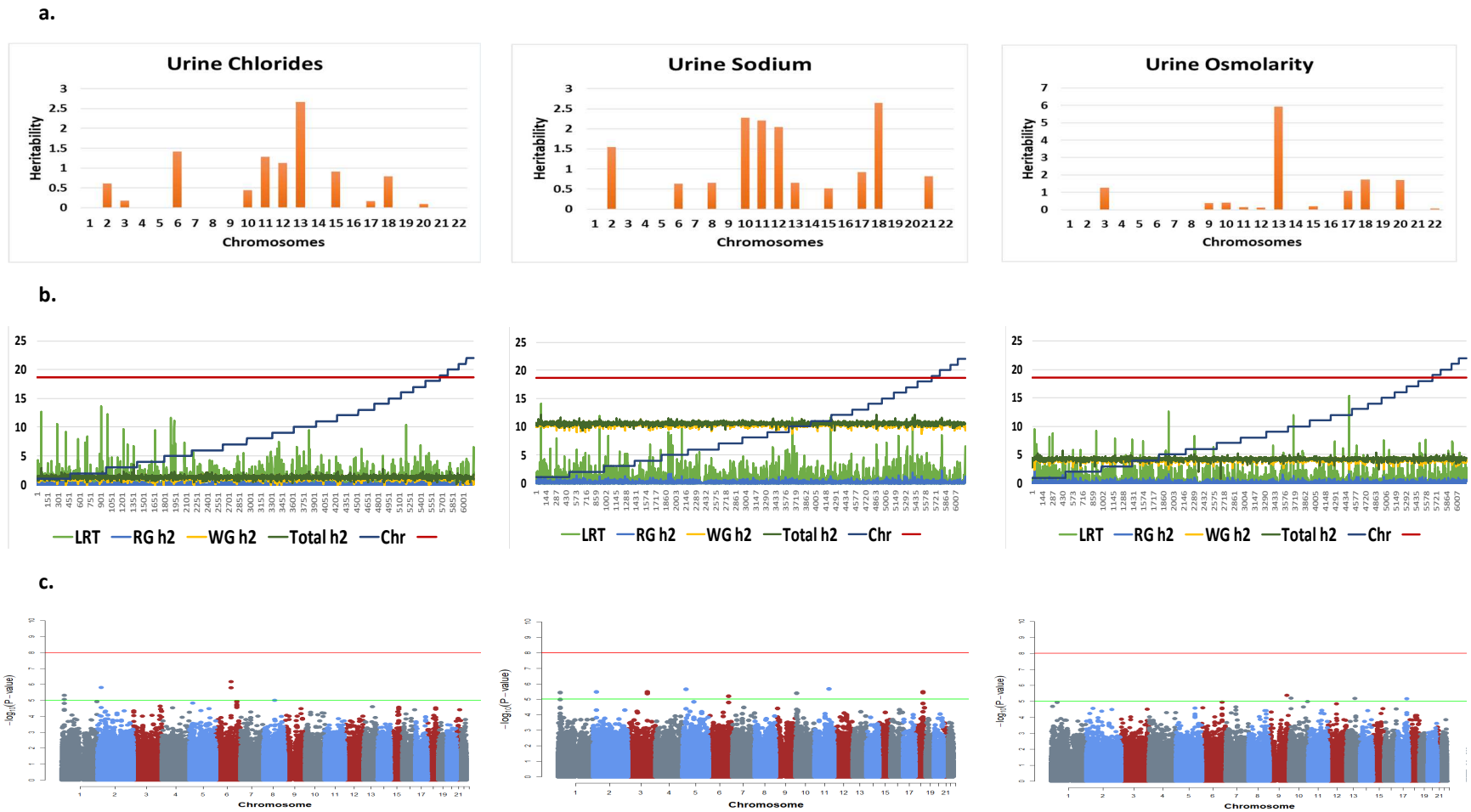
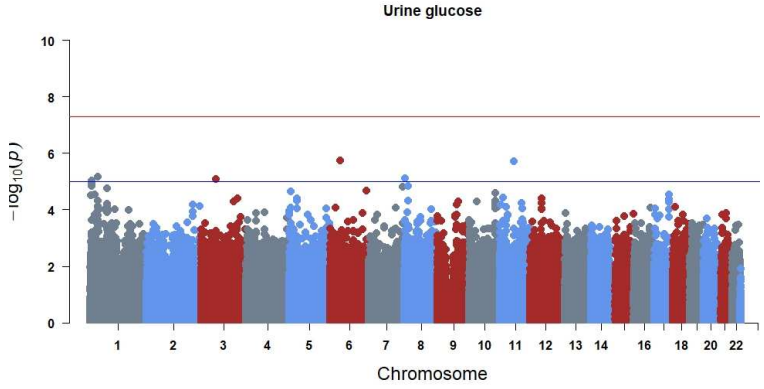
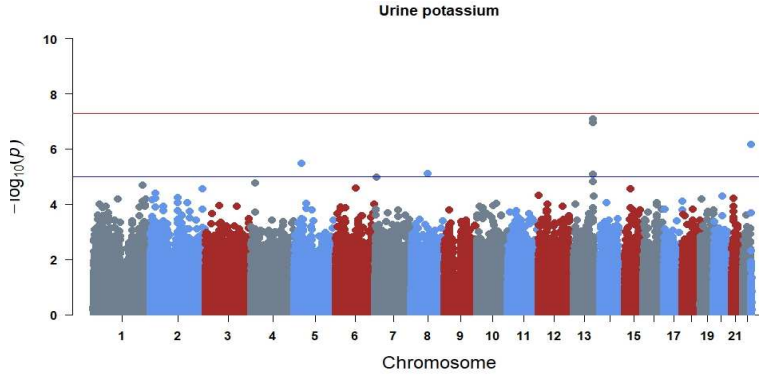



Figure 2.5. Whole genome and regional analysis of urine chlorides, sodium and osmolarity. **a** is plots chromosomal heritability, with the heritability of chromosomes expressed as percentage**. b** is plots regional of GREML analysis. The Likelihood ratio test statistic (LRT), regional heritability (RG h2), heritability of rest of genome excluding region (WG h2) and total heritability (Total h2) are plotted. The red horizontal line is Bonferroni-corrected threshold for genome-wide significance (LRT=18.6, p-value < 1.6278e-05). Chromosomes are indicated by blue steps plot. **c** is the Manhattan plots showing genome-wide associations. The points are plots of -log10 of association p-values of genome-wide SNPs. Red line is genome-wide significance (GWS) threshold (p-value < 1e-08) and green line is threshold for significance at the suggestive level (p-value < 1e-05). For all the three traits, the Manhattan plot shows association at the suggestive level as no SNP reaches genome-wide significance

a.




b.




c.

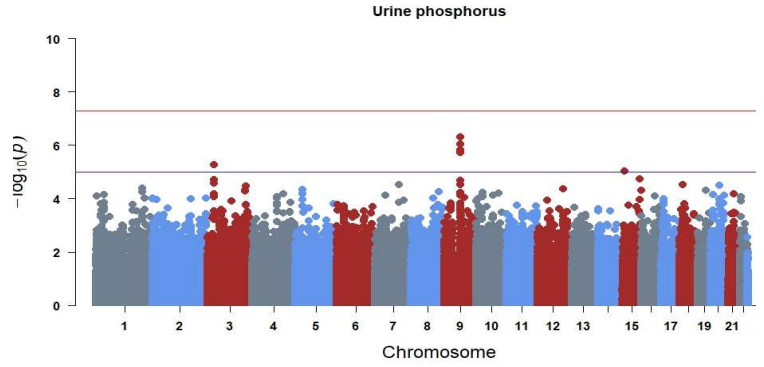



Figure 2.6. The genome-wide evidence for SNP association for urine glucose, potassium and phosphorus. The points are plots of -log10 of association p-values of genome-wide SNPs. Red line is genome-wide significance (GWS) threshold and blue line is threshold for significance at the suggestive level (p-value < 1e-05). For all the three traits, the Manhattan plot shows association at the suggestive level as no SNP reaches genome-wide significance.

Table 2.3. The markers most strongly associated with the 8 urine traits analysed using a mixed model GWAS for the trait transformed GS: SFHS dataset. The trait name, genomic inflation factor, top SNPs, chromosomal position of SNPs, the effect of the minor allele, p-value corrected for genomic inflation, and genes within 1MB of SNP.

| Traits | λ | SNP | Chr. | effect | P | Genes within 1MB |
|---|---|---|---|---|---|---|
| Calcium | 1.000 | rs4574243 | 3 | 0.043 | 9.3e-09 | RBMS3 |
| | | rs6934483 | 6 | 0.033 | 5.9e-06 | HULC, SLC35B3, LOC100506207 |
| Chlorides | 1.005 | rs12197896 | 6 | -0.428 | 7.2e-07 | - |
| | | rs7567950* | 2 | 0.661 | 1.6e-06 | - |
| | | rs11588628 | 1 | 0.053 | 5.0e-06 | KAZN, TMEM51, FHAD1, CELA2A, AGMAT, C1orf126, EFHD2, DNAJC16, CTRC, CELA2B, CASP9 |
| | | rs3804538 | 6 | 0.273 | 1.2e-05 | ADAT2, LOC285740, PEX3, FUCA2, PHACTR2, LTV1, ZC2HC1B, AIG1, PLAGL1, LOC100507489 |
| Glucose | <1 | rs3756804 | 6 | 0.196 | 1.8e-06 | CLIC5, ENPP4, ENPP5, RCAN2, CYP39A1 |
| | | rs12293191 | 11 | 0.121 | 1.9e-06 | CCDC86, ZP1, PRPF19, PTGDR2, TMEM109, TMEM132A, CD6, CD5, SLC15A3, VPS37C, PGA3 |
| Potassium | <1 | rs9518824 | 13 | 0.079 | 8.3e-08 | CCDC168, TPP2, METTL21C, TEX30, KDELC1, ERC5, BIVM-ERCC5, METTL21EP |
| | | rs3797064 | 5 | 0.098 | 3.4e-06 | PDZD2, GOLPH3, MTMR12, ZFR, MIR4279 |
| | | rs1796120 | 7 | -0.079 | 1.0e-05 | C1GALT1, COL28A1, LOC100131257, MIOS, RPA3, LOC729852, RSPH10B2, RSPH10B, CCZ1B |
| Magnesium | <1 | rs2541872** | 18 | 0.045 | 1.4e-06 | CDH7 |
| | | rs4606726** | 16 | -0.043 | 2.2e-06 | GP2, UMOD, PDILT, ACSM2A, ACSM2B, ACSM1, THUMPD1, ACSM3, ERI2, LOC81691 |
| | | rs16837959** | 2 | -0.055 | 8.3e-06 | TMEM237, MPP4, ALS2, ALS2CR11, CDK15, STRADB, TRAK2, ALS2CR12, CASP8, CASP10 |
| Sodium | 1.000 | rs12365770 | 11 | 0.174 | 2.1e-06 | CNTN5 |
| | | rs7567950* | 2 | 0.447 | 3.3e-06 | - |
| | | rs12634170 | 3 | -0.179 | 3.3e-06 | NMNAT3, RBP1, RBP2, COPB2, MRPS22, CLSTN2, PISRT1 |
| | | rs12604137 | 18 | 0.170 | 3.6e-06 | CCBE1, RAX, CPLX4, LMAN1, GRP, SEC11C, OACYLP, ZNF532, PMAIP1 |

Table continues

| | | | | | | |
|---|---|---|---|---|---|---|
| | | rs8161 | 6 | 0.177 | 6.3e-06 | ADAT2, PEX3, FUCA2, LOC285740, PHACTR2, LTV1, ZC2HC1B, PLAGL1, AIG1, LOC100507489 |
| Osmolarity | 1.001 | rs12551209** | 9 | 1.342 | 4.2e-06 | GGTA1P, DAB2IP, STOM, GSN, RAB14, CNTRL, C5, TTLL11 |
| | | rs12782067 | 10 | 0.986 | 6.4e-06 | UPF2, DHTKD1, SEC61A2, CDC123, NUDT5, PROSER2, PROSER2-AS1, ECHDC3, USP6NL |
| | | rs8001997 | 13 | 1.042 | 6.7e-06 | MYCBP2, MYCBP2-AS1, FBXL3, CLN5, IRG1, BTF3P11, KCTD12, SCEL, MIR3665, SLAIN1 |
| | | rs4147996* | 17 | 2.117 | 6.8e-06 | - |
| Phosphorus | <1 | rs1414850** | 9 | 0.044 | 4.9e-07 | KLF9, TRPM3, SMC5, MIR204, LOC100507299, LOC100507244, MAMDC2 |
| | | rs7623722** | 3 | -0.028 | 5.3e-06 | LINC00693, RNMS3, ZCWPW2, AZI2 |

* Rare SNPs (1% < MAF < 5%)

** SNP association disappears in untransformed data

Table 2.4. The markers most strongly associated with the raw measurement of urine traits analysed using a mixed model GWAS for the GS: SFHS dataset. The trait name, genomic inflation factor, top SNPs, chromosomal position of SNPs, the effect of the minor allele, p-value corrected for genomic inflation, and genes within 1MB of SNP.

| Traits | λ | SNPs | Chr | effect | P | Genes within 1MB |
|---|---|---|---|---|---|---|
| Calcium | <1 | rs4574243 | 3 | 0.725 | 1.7e-06 | RBMS3 |
| | | rs12341976** | 9 | -0.679 | 7.6e-06 | PBX3, MAPKAP1, GAPVD1 |
| | | rs4742914** | 9 | 0.668 | 1.8e-05 | NIPSNAP3A, NIPSNAP3B, LOC286367, ABCA1, SMC2, OR13 Family of genes |
| Chlorides | 1.005 | rs7567950* | 2 | 43.158 | 1.6e-06 | - |
| | | rs12197896 | 6 | -26.22 | 3.3e-06 | - |
| Glucose | 1.015 | rs11597250** | 10 | 317.26 | 2.9e-11 | DOCK1, C10orf90, LINC00601, FAM196A, ADAM12 |
| | | rs7007992** | 8 | 378.17 | 2.9e-11 | PAG1, FABP5, ZNF704, PMP2, FABP9, FABP4, FABP12 |
| | | rs960491* | 19 | 314.02 | 4.4e-11 | DOT1L, LINGO3, PLEKHJ1, IZUMO4, AP3D1, SF3A2, AMH, MIR4321, JSRP1, OAZ1 |
| | | rs17013067** | 4 | 270.18 | 2.7e-10 | IBSP, MEPE, HSP90AB3P, SPP1, PKD2, DMP1, DSPP, SPARCL1, NUDT9, ABCG2, HSD17B11 |
| | | rs2389225** | 13 | 211.35 | 4.3e-08 | ABCC4, CLDN10, CLDN10-AS1, DZIP1, DNAJC3 |

Table continues

| | | | | | | |
|---|---|---|---|---|---|---|
| Potassium | <1 | rs9518824 | 13 | 12.58 | 4.8e-07 | CCDC168, TPP2, METTL21C, TEX30, KDELC1, ERC5, BIVM-ERCC5, METTL21EP |
| | | rs1348259** | 5 | 15.155 | 4.1e-06 | LINC00052, NTRK3, AGBL1 |
| | | rs1796120 | 7 | -13.38 | 1.1e-05 | C1GALT1, COL28A1, LOC100131257, MIOS, RPA3, LOC729852, RSPH10B2, RSPH10B, CCZ1B |
| Magnesium | <1 | rs1441827** | 15 | 0.736 | 1.4e-06 | ALDH1A2, AQP9, MYZAP, GCOM1, POLR2M, CGNL1 |
| | | rs6953804** | 7 | -0.470 | 8.8e-06 | CARD11, SDK1, GNA12, AMZ1, TTYH3, IQCE, LFNG, MIR4648, BRAT1 |
| Sodium | 1.00 | rs7567950* | 2 | 30.98 | 7.9e-07 | - |
| | | rs12365770 | 11 | 11.688 | 1.0e-06 | CNTN5 |
| | | rs12634170 | 3 | -11.715 | 3.0e-06 | NMNAT3, RBP1, RBP2, COPB2, MRPS22, CLSTN2, PISRT1 |
| | | rs8161 | 6 | 11.704 | 4.6e-06 | ADAT2, PEX3, FUCA2, LOC285740, PHACTR2, LTV1, ZC2HC1B, PLAGL1, AIG1, LOC100507489 |
| | | rs12604137 | 18 | 10.891 | 5.2e-06 | CCBE1, RAX, CPLX4, LMAN1, GRP, SEC11C, OACYLP, ZNF532, PMAIP1 |
| Osmolarity | 1.00 | rs8001997 | 13 | 16.071 | 2.5e-06 | MYCBP2, MYCBP2-AS1, FBXL3, CLN5, IRG1, BTF3P11, KCTD12, SCEL, MIR3665, SLAIN1 |
| | | rs4147996* | 17 | 32.547 | 2.7e-06 | - |
| | | rs699608 | 12 | 15.365 | 4.1e-06 | PPM1H, AVPR1A, C12orf61, MIRLET7I, MON2 |
| | | rs12782067 | 10 | 14.516 | 6.7e-06 | UPF2, DHTKD1, SEC61A2, CDC123, NUDT5, PROSER2, PROSER2-AS1, ECHDC3, USP6NL |
| Phosphorus | 1.004 | rs9468226** | 6 | 17.749 | 7.9e-07 | LOC100507173, HIST1H2AL, LOC100131289, HIST1H2BL, OR2B2, HIST1H2AI, ZNF184 |
| | | rs3980449** | 11 | 7.578 | 9.8e-07 | PAMR1, FJX1, TRIM44, SLC1A2, CD44, LDLRAD3, MIR3973 |
| | | rs10508250** | 10 | 10.443 | 1.8e-06 | TRPM3, KLF9, SMC5, MIR204, SMC5, LOC100507299, LOC100507244, MAMDC2 |
| | | rs9933070 | 16 | 9.914 | 1.9e-06 | XYLT1 |

* Rare SNPs (1% < MAF < 5%)

** SNP association disappears in transformed data

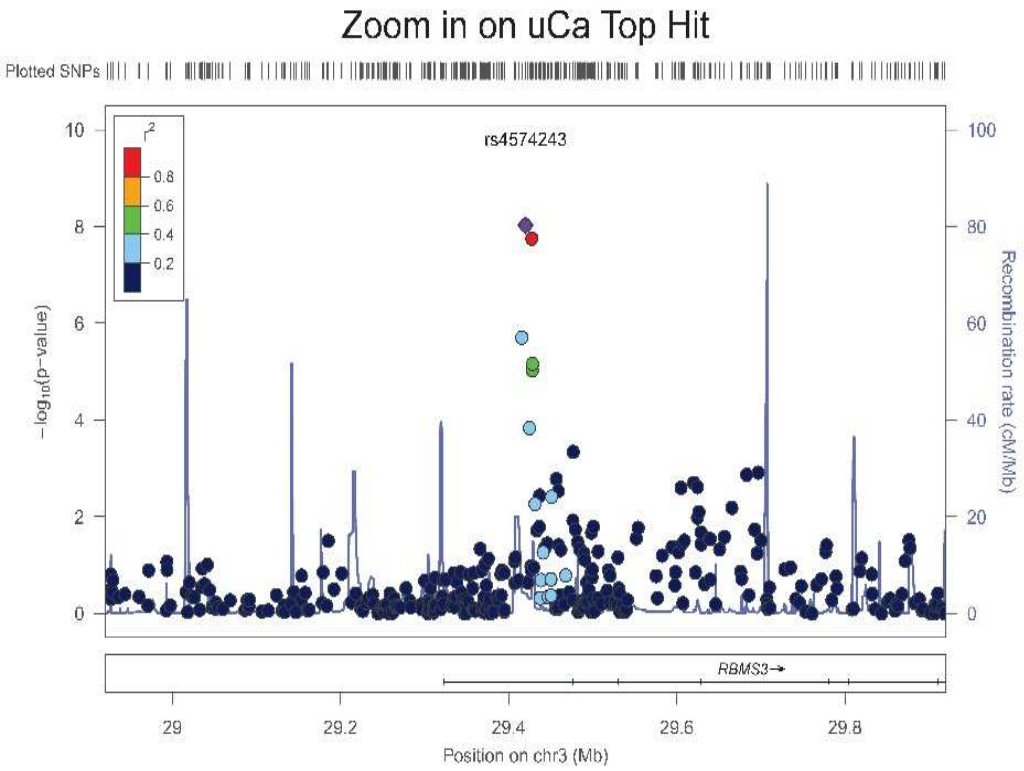



Figure 2.7. LocusZoom plot around the top hit SNP for urine calcium on chromosome 3. The plot shows that the two highly associated variants are strongly correlated in terms of LD (LD > 0.8). The top SNP rs4574243 lies in the RBMS3 (RNA Binding Motif, Single Stranded Interacting Protein 3) gene is implicated in DNA replication, gene transcription, cell cycle progression and apoptosis. Diseases associated with RBMS3 are osteonecrosis of the jaw and osteoporosis.

## 2.4 **Discussion**

The genome-wide association approach to studying complex traits has taken centre stage in genetics research during the last decade (Björkegren et al., 2015; Burton et al., 2007; McCarthy et al., 2008; Seng and Seng, 2008) and has had an enormous impact particularly on the field of human genetics (Bush and Moore, 2012). Despite the number of problems catalogued on GWAS, such as multiple testing, low power to detect effects and the winner's curse (Shriner et al., 2007; Wray et al., 2013), the approach still enjoys much usage in the study of genetic control of

complex traits (Welter et al., 2014) which only highlights the fact that the GWA approach has come to stay. This study, therefore, aimed to investigate, using the GWA approach, if the kidney phenotype measurements taken for individuals of the GS: SFHS are under significant genetic control.

Before attempting to answer that question, the ancestry of the study individuals was checked by contrasting them against the background of the several ancestral populations from the 1000 Genomes Project. The results from the principal components (PC) analysis shown in Figure 2.1b show that the individuals analysed are of European ancestry since they clustered together with 1000 genomes populations from Europe. A further investigation of the GS: SFHS population alone (Figure 2.1a) to check for genetic substructure showed seven individuals clustering away from the rest of the population on the second principal component axis. The pedigree information showed that they are individuals of the same family spanning three generations from grandparents to grandchild. Those individuals were kept in downstream mixed-model GWA analysis which fitted a GRM to account for any potential confounding due to the genetic substructure. The genomic inflation parameters (lambda) reported for the traits in Table 2.3 and Table 2.4 were all 1 or acceptably close to 1. This suggests there is no strong influence of genetic substructure or other design factors on the association test statistic (Aulchenko, 2011).

The GWA analysis results for the kidney phenotypes studied show some associations of SNPs at the suggestive level with one SNP (rs4574243) on chromosome 3 reaching genome-wide (GW) significance (p-value $< 9.378709 \times$

$10^{-9}$) for urine calcium in the transformed data (Table 2.3 and Figure 2.4a). A follow-up analysis on the top GWAS hits for these traits revealed that most of them lie in regions of the genome that have several genes nearby. However, the functional relevance of these genes to the phenotypes of interest remains to be explored. A quick look up, nonetheless, for the RBMS3 gene found in the region of the SNP that reached genome-wide significance for urine calcium (Figure 2.4a), showed that the gene has been implicated in osteonecrosis of the jaw (Nicoletti et al., 2012), and also reported to interact with ZNF516 to influence bone mineral density (Yang et al., 2013). A SNP in the RBMS3 gene region was found to be associated with trochanter bone mineral density in males in the Framingham heart study (Kiel et al., 2007). This link of RMBS3 with osteoporosis phenotypes might explain its association to urine calcium concentration.

The GWA analysis for the untransformed data saw 4 SNPs for urine glucose reaching genome-wide significance (Table 2.4). All these SNPs were rare (1% < MAF < 5%) and the signal disappeared in the transformed data. The SNP effect sizes in the transformed data were greatly reduced (Table 2.3). Both the Box-Cox and log10 transformations involve raising the values of the data points to a certain exponent and this will change the scale of the data which invariably will change the scale of the effect sizes. However, how much information is lost by this change is not known. For most of the traits, the significance of the SNP association is improved by data transformation because transformation improved model fit.

Top hit SNPs that were rare (1% < MAF < 5%) had relatively bigger effect sizes than the common SNPs (Table 2.3 and Table 2.4). Natural selection will drive the

frequencies of variants with relatively large effects on disease risk down (Bodmer and Bonilla, 2008). The consequence of this low SNP MAFs is that such rare variants may be missed in GWA studies because they will fail to pass the commonly used MAF threshold (5%) set in initial QCs. Again, for studies like this which include rare variants in GWA analysis, even if they show up in association signals, they will most likely be overlooked in follow-up studies. The reason being that the GWA field is chiefly driven by the 'common disease – common variant' hypothesis (Cirulli and Goldstein, 2010) and also the lack of power to replicate rare associations. Although rare variants could have a prominent claim as key drivers of some common diseases (Bodmer and Bonilla, 2008; Pritchard, 2001). In spite of this, variants having large effect sizes doesn't guarantee any functional relevance since association doesn't necessarily imply causation. This was the case for most of the rare variants identified in Table 2.3 and Table 2.4, which had no genes within 1MB of those GWAS hits. I must admit, however, that the further analysis employed in this study to search for genes within 1MB genomic regions of the GWAS hits is not an exhaustive solution since the variants could be exerting their influence on genes or causal variants several megabase-pairs away in the genome. Also, not all causal variants have to be genes; regulatory elements that can affect phenotypes can be far away from genes.

The modest heritability estimates obtained from the GREML analysis using GCTA (Table 2.2), suggest that the kidney phenotypes may not be under a significant genetic control of common SNP variants. The standard errors of most of the heritability estimates show that they are not significantly different from zero. But this initial assertion would require further analysis with a larger sample size to confirm it

since standard error decreases with sample size. The local chromosomal estimates of heritability were calculated and summed up for each phenotype and all the traits curiously had a significant increase in the heritability estimates (Table 2.2). This could be due to the sum of the local chromosomal estimates of variance being driven up by the covariance between the chromosomes. This covariance between chromosomes may not be driven by LD since LD may be limited between SNPs on different chromosomes. But in a data with lots of family structure like the GS: SFHS, individual chromosomes will each reflect the family structure to some extent. So, the family component of the heritability may be included in the individual chromosomal heritabilities and so may be counted multiple times when you sum their effects.

The results for the regional heritability analysis for most of the traits showed some consistency with the chromosomal heritability results and with the GWAS results. This kind of analysis improves the resolution of the heritability down to the very few SNPs in the regions that may be driving the underlying genetic variance. It can, therefore, give evidence in support of some of the associations found in the GWA analysis of traits. For instance, the plots shown in Figure 2.4 and 2.5, for the urine traits, show that regions of top GWAS hits have relatively high estimates of local heritability compared to other genomic regions. For instance, in calcium, the Manhattan plot shows an association peak on chromosome 3 and the regional heritability plot also shows the most significant region on chromosome 3 (LRT = 16.553, p-value $2.37 \times 10^{-5}$), which is just shy of genome-wide significance (p-value $1.6278 \times 10^{-5}$) after multiple testing correction for the 6143 regions tested in the regional heritability analysis. The plot from the chromosomal heritability analysis for

urine calcium also points to chromosome 3 as having a local heritability of 4.65% (SE = 4.07) explaining nearly a quarter (22.43%) of the total chromosomal heritability 20.74% (SE = 7).

In summary, mixed model GWAS has allowed us to scrutinize kidney disease endophenotypes in the GS: SFHS data, expanding on the myriad of GWA studies on complex traits that have resulted since the first GWAS a decade ago. There have been very sparse GWAS reports on these kidney disease endophenotypes which make these findings uniquely important. Many of the GWAS hits for these traits lie within areas of the genome that have genes close by which presents a useful avenue to be explored in future studies. Data transformation improved evidence of SNP associations in most traits. The heritability estimates calculated from the analysis suggest that these kidney disease endophenotypes are under modest genetic control common genetic variants. The study would have benefitted from larger sample size because that would have increased the power to detect more SNP associations and reduce the standard errors of the heritability estimates. However, for now, understanding the implications of these findings in the context of kidney disease development should offer some useful leads in the fight against the disease.

# Chapter 3

## 3 Use of a Bayesian mixture model (Bayes R) to investigate the genetic control of complex traits

### 3.1 Introduction

The genome-wide association (GWA) analysis of complex traits has made a lot of advances in human genetics and has been employed to estimate SNP heritability and map genetic loci of interest in many traits. The approach also offers itself to the investigation of the underlying genetic architecture of traits, and to predict the genetic or breeding value of individuals (Gianola, 2013; Moser et al., 2015) through the polygenic risk score analysis (Wray et al., 2007).

The general framework for the GWA analysis is to regress the measured responses or phenotypes of study individuals on whole-genome markers' codes, mostly derived from genotyped SNPs of those individuals, to determine the association between SNPs and a trait. From a frequentist standpoint, evidence of an association between phenotypes and genotyped SNPs is assessed with p-values calculated for the null hypothesis of no association (Stephens and Balding, 2009).

The conventional GWA model treats the effects of the markers as fixed and does not analyse them simultaneously. Thus, at each instance of marker effect estimation, the model ignores the effects of all other markers in the background. This analysis of markers in isolation often leads to an increase in the residual variance and a decrease in the power to capture true associations (de los Campos et al., 2010; Hoggart et al., 2008).

Other factors such as the minor allele frequency (MAF) and the number of participants in the study also affect the power of the test for association (Stephens and Balding, 2009). SNPs with very low MAF are considered rare and are usually removed from GWA analysis during data quality control to improve power and reduce the risk of finding spurious associations.

There is also the problem of multiple testing which arises statistically due to the test for SNP associations to phenotypes not being done simultaneously. Stringent p-value thresholds are set to account for the problem of multiple testing in GWA analysis which in itself presents the risk of true associations being discarded while overestimating the effects of SNPs that pass the stringent threshold.

The overestimation of the SNP effects is due to the so-called “winner’s curse” or Beavis effect (Beavis, 1997). The winner’s curse arises in GWA analysis because the analysis involves numerous markers (mostly in the hundreds of thousands), which generally have very small effects. Consequently, there is a low power to detect markers that are significant, thus the effects of those markers that come out as significant tend to be overestimated.

One direct effect of the winner's curse in GWA studies is the non-reproducibility of GWAS hits in studies that aim to replicate GWAS results. In addition, in most complex traits, the markers that pass the significance threshold explain just a fraction of the heritability, and this led to claims that a proportion of the heritability was missing (Maher, 2008) for most traits. There has since been evidence to suggest that the heritability is not missing (Yang et al., 2015) but perhaps only hidden because of an even lower power to detect associations of markers of small effects (Gibson, 2010; Park et al., 2010).

From what has been said thus far, it can be argued that models that treat the SNP effects as random and analyse them all simultaneously (Meuwissen et al., 2001; VanRaden, 2008; Yang et al., 2010) should perform better than the conventional approach of treating the effects as fixed and analysing SNPs one after the other. Thus models that treat the SNP effect as random have been developed and these models fit a genomic relationship matrix (GRM) to account for all SNP effects simultaneously (Fernando and Grossman, 1989; Habier and Fernando, 2013; VanRaden, 2008). One of such methods is the Genomic Best Linear Unbiased Predictor (GBLUP), which draws its theoretical foundations from an equivalent model (BLUP) that is utilised by animal breeders. GBLUP fits a GRM in place of the pedigree-based numerator relationship matrix used in BLUP. The GRM is generated by scaling the covariance of genomic values between two individuals to be equivalent to the pedigree-based elements of the numerator relationship matrix (VanRaden, 2008). Two scaling factors were developed by VanRaden (2008) but several scaling factors have since been developed (Wang et al., 2017). Other methods that utilise the GRM in a genome-wide

analysis setting involve the use of a restricted maximum likelihood estimation of genetic and residual components of the phenotypic variation (Speed et al., 2012; Yang et al., 2010).

The GBLUP model assumes that all SNPs have small effects and these effects are drawn from the same normal distribution (Strandén and Garrick, 2009; VanRaden, 2008). This essentially means that all SNPs have an equal chance of contributing to the trait variation. This assumption suffers setbacks in at least two ways. First, it is not prudent to blindly assume *a priori* that all genetic loci will be associated with a trait and thus should have an effect. This is because, in truth, not all loci will be associated with a trait or will be in LD with the true underlying causal loci for a trait. Second, some quantitative traits have been shown to have a few loci with moderate to large effects on those traits, which means those loci can't be treated the same as every other locus. There is, therefore, some benefit to be gained by fitting models that adequately account for the background effects of polygenic loci as well as the loci with moderate to large effects.

After dealing with all problems described above such as low power to detect effects, multiple testing and the winner's curse, there is one last hurdle to overcome in a GWA analysis. And that is GWA models typically having more predictors, in most cases hundreds of thousands of markers, than phenotype records, which introduces the problem of high dimensionality.

Bayesian models have been presented as being better equipped than GBLUP to adequately deal with the challenges that conventional GWA analysis face.

Bayesian models allow for SNP effects to be sampled from more than one distribution which eliminates the problem of assuming that all genetic loci will have an effect on a trait. They also deal with the problem of dimensionality by introducing sparseness into the model. They do this by either shrinking the effects of those SNPs with very small effects to zero as in Bayesian LASSO (Tibshirani, 1994) or by choosing a subset of the markers to have an effect. Meuwissen et al. (2001) advanced the second approach by proposing Bayesian methods, BayesA and BayesB which sample SNP effects from prior distributions that accounts for the different SNP effect sizes. These methods fit a hierarchical model that samples the genetic variances of the markers from an inverse chi-squared conditional prior distribution (Meuwissen et al., 2001). BayesA and BayesB led to the proliferation of other Bayesian methods over the years each unique in the prior distribution it samples the SNP effects from in the GWA analysis. Gianola et al. (2009) came up with the term 'Bayesian alphabet' to refer to the growing number of Bayesian methods employed in GWA analysis named after letters of the English alphabet (Gianola, 2013).

An extension to the BayesB model was proposed by Erbe et al. (2012) to improve the sampling of SNPs with zero and nonzero effects, that is which SNP is in or out of the subset of SNPs chosen to be having an effect on the trait. They proposed a hierarchical model that sampled the SNP effects from a mixture of 4 normal distributions, one with zero contribution to the total genetic variance and the other three with an increasing contribution to the genetic variance. In keeping with the tradition of the Bayesian alphabets, the new method was named BayesR.

In this chapter, I explore the BayesR method in some detail and apply it to simulated phenotypes in a simulation study to assess the performance of this BayesR method as compared to that of GBLUP. I also apply this model in the subsequent chapter on urine phenotypes measured from individuals from the Scottish population. The underlying genetic architecture of these traits was investigated using the method and an estimate of the additive genetic variance explained by the genome-wide SNPs was obtained.

## 3.2 Methods

### 3.2.1 Mixture models in the GWA setting

Mixture models are useful in determining the inner structure of data with hidden clusters when no information is available on how the clusters might have been formed (Picard, 2007). These models are composed of a finite or infinite number of components that can describe different attributes of data (Marin et al., 2005). Each component may or may not come from the same type of distribution.

In a GWA setting, the different components of the model can describe different classes of markers that have different effect sizes on traits. Consider a vector $\boldsymbol{g}$ containing the allelic substitution effects of $p$ genotyped SNPs for a measured phenotype. In a mixture model context, the density function for the realisation of $\boldsymbol{g}$ given all unknown parameters $\boldsymbol{f}(\boldsymbol{g}|\boldsymbol{\varphi})$ is assumed to be a mixture of $K$ parametric distributions such that

$$f(g|\varphi) = \sum_{k=1}^{K} \pi_k f(g|\theta_k), \qquad \sum_{k=1}^{K} \pi_k = 1, \qquad K > 1, \tag{3.1}$$

where $\boldsymbol{\varphi}$ is a vector of all unknown parameters of the model, $\boldsymbol{\pi_k}$ is the mixing proportion or the weight of component $k$. If all the $f(g|\theta_k)$ are normal distributions, then $\theta$ represents the unknown mean and variance.

Primarily, such models aim to recover unobserved clusters in the data by attempting to label each data point as belonging to a particular cluster. In a GWA study, these could be two-cluster labels of whether a locus has an effect on a trait or not. This clustering on locus effects can be more than two such as a locus having no effect, a small effect, or a large effect. If a locus is classified as belonging to one component of the mixture distributions, then it will be classified as zero for all others. If these class labels were known, then the estimation of the mixture parameters would be easy since the parameters of each mixture distribution $f(g|\theta_k)$ could be estimated directly from the data points from a component $k$ (Picard, 2007). The labels are unobserved; however, and thus estimation of mixture parameters can only be based on the observed data points.

Fitting mixture distributions can be done using a number of approaches, such as graphical methods, the method of moments, maximum likelihood and Bayesian approaches (Picard, 2007). Bayesian inference offers the opportunity for probability statements to be directly made about the unknown parameters (i.e. the mixing proportions, the mean, the additive genetic variance and error variance), and the priors used in the analysis (Marin et al., 2005). BayesR fits a mixture of four normal

distributions and implements a Gibbs sampler that iteratively estimates the model parameters.

### 3.2.2 The Bayesian mixture model

BayesR was developed by Erbe et al. (2012). The BayesR model starts with the general setting similar to the one described in the equation 3.1 above and assumes a mixture of four normal distributions with means equal to zero and variances that account for increasing proportions of the additive genetic variance as the conditional priors for the distribution of SNP effects

$$p\left(g_j\middle|\pi_k\sigma_g^2\right) = \pi_1 N\left(0, 0\times\sigma_g^2\right) + \pi_2 N\left(0, 10^{-4}\times\sigma_g^2\right) + \pi_3 N\left(0, 10^{-3}\times\sigma_g^2\right) + \pi_4 N(0, 10^{-2}\times\sigma_4^2). \quad (3.2)$$

Here, $\pi_k$ are the mixing proportions and $\sigma_g^2$ is the additive genetic variance explained by all the SNPs. Thus, under this model, if a SNP effect is drawn from the first component of the mixture, it will be zero with a zero variance. This incorporates sparseness into the model since a large proportion of the SNPs are expected to have no effect on traits and thus go into this distribution. SNP effects drawn from the second component will be normally distributed with mean zero and variance = $10^{-4}\times\sigma_g^2$.

In the implementation of the model, uninformative flat priors for the mixing proportions are generally assumed (Erbe et al., 2012), although the model can accommodate informative priors. The prior distribution of the proportions of SNP in each component is a Dirichlet distribution, with parameter vector $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3, \alpha_4) = 1$. This defines a uniform prior for the mixture distributions and

thus each SNP will have an equal chance of being in any of the distributions. The posterior distribution of the mixing proportions is $\boldsymbol{\pi_k} \sim Dir(\boldsymbol{\alpha} + \boldsymbol{\beta})$**,** where $\boldsymbol{\beta}$ is a vector containing the number of SNP in each mixture distribution estimated from the data (Erbe et al., 2012). The $\boldsymbol{\beta}$ estimates are obtained by first calculating four likelihoods assuming a particular SNP is in one of the four mixture distributions at a time with a corresponding probability (Erbe et al., 2012). The mixture distribution to sample the SNP effect from is then selected based on these probabilities for the current iteration (Erbe et al., 2012).

The genetic variance $\sigma_g^2$ and the error variance $\sigma_e^2$ have a starting value of 0.01. The priors for both $\sigma_g^2$ and $\sigma_e^2$ are assumed to be from a scaled inverse chi-squared distribution parameterised by the number of chi-squared degrees of freedom $(\nu_0)$ and the scaling parameter$(\tau_0^2)$. The priors used for the two hyper-parameters for the scaled inverse chi-squared distribution are -2 and 0 respectively.

Updated estimates of the pre-specified parameters ($\sigma_g^2$, $\pi_k$, $\sigma_e^2$), and an estimate of the other unknown parameters, the overall mean $\mu$ and the SNP effects $g$ are estimated using a Gibbs sampling scheme that samples values from each parameter's conditional posterior distribution (conditional on other parameters). BayesR uses a Gibbs sampling method analogous to the scheme described for BayesA (Meuwissen et al., 2001) with an addition of a polygenic effect, and with the SNP effects coming from four normal distributions with different variances (Erbe et al., 2012).

### *3.2.2.1 The Gibbs sampler used*

My implementation of the BayesR model was performed with a Fortran program written by Moser et al. (2015) and compiled locally using the Intel Fortran compiler. The Gibbs sampler used as described in the supplementary text of Moser et al. (2015) is given as follows:

The sampler first samples the overall mean from the full conditional posterior distribution

$$\mu|. \sim N\left(n^{-1}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}g_j\right), \frac{\sigma_e^2}{n}\right) \quad (3.3)$$

where $y_i$ is the observed phenotype for individual $i$, $n$ is the total number of individuals, $\mu$ is the overall mean, $x_{ij}$ is the number of copies of the effect allele of SNP $j$ which takes the values 0, 1, or 2, and $g_j$ the allelic substitution effect of SNP $j$. The starting values for the $g$ and $\sigma_e^2$ are 0 and 0.01 respectively.

It then calculates the probability that a SNP $j$ is in component $k$ of the mixture distribution. The log likelihood of SNP $j$ being in component $k$ is

$$LogL(j,k) = \log(\pi_k) - \frac{1}{2\sigma_e^2}\left(\sum_{i=1}^{n}\tilde{y}_i^2 - \mu_{j,k}\sum_{i=1}^{n}\left(x_{ij}\tilde{y}_i\right)^2\right) - \frac{1}{2}logV, \quad (3.4)$$

where $\tilde{y}_i$ is the phenotype of individual $i$ corrected for the overall mean and the effects of all markers in the model, except marker $j$.

$$\tilde{y}_i = \left(y_i - \mu - \sum_{l\neq j}^{p} x_{il}g_l\right)$$

and

$$\mu_{jk} = \frac{\sum_{i=1}^{n} x_{ij}\tilde{y}_i}{\frac{\sum_{i=1}^{n} x_{ij}^2 + \sigma_e^2}{\sigma_k^2}}$$

logV is the log likelihood of the reduced model including only the effect of the marker $j$ and a residual effect

$$logV = \left| nlog(\sigma_e^2) + \log\left(\frac{\sigma_k^2}{\sigma_e^2}\sum_{i=1}^{n} x_{ij}^2 + 1\right)\right|. \quad (3.5)$$

Then the probability of marker $j$ being in component $k$ is

$$\Pr(x_{j\in k}) = 1\Big/\sum_{k=1}^{K} \exp[L(i,l) - L(i,k)]\,. \quad (3.6)$$

The sampler assigns marker $j$ to component $k$ based on a value sampled from a uniform distribution.

The third step is to sample the allelic substitution effect for marker $j$ from mixture component $k$ from the full conditional posterior distribution.

$$g_{jk}|. \sim N(\mu_{jk}, S_k^2),$$

where

$$S_k^2 = \frac{\sigma_e^2}{\sum_{i=1}^{n} x_{ij}^2 + \sigma_e^2/\sigma_k^2}$$

and $\mu_{jk}$ is sampled as above.

The sampler repeats for all genotyped SNPs, the steps of calculating probability and assigning a marker to a mixture component and then sampling the marker effects from the assigned mixture component.

After assigning all markers to a mixture distribution, the additive genetic variance is then sampled from the full conditional posterior distribution

$$\sigma_g^2|. \sim Inv - \chi^2\left(\nu_0 + m_g, \frac{m_g \sum_{j=1}^{p} g_j^2 + \nu_0 \tau_0^2}{\nu_0 + m_g}\right), \tag{3.7}$$

where $m_g$ is the number of SNPs included in the current model, $\nu_0$ is the number of chi-squared degrees of freedom and $\tau_0^2$ is the scaling parameter of the scaled inverse chi-squared distribution. These parameters have values of -2 and 0 respectively. The residual variance is also inverse chi-square distributed and is sampled from the full conditional posterior distribution

$$\sigma_e^2|. \sim Inv - \chi^2\left(\upsilon_0 + n, \frac{\sum_{i=1}^{n}(y_i - \mu - \sum_{j=1}^{p} x_{ij} g_j)^2 + \upsilon_0 \tau_0^2}{\upsilon_0 + n}\right). \tag{3.8}$$

The sampler then updates the mixing proportions by sampling from the posterior

$$\pi|. \sim Dirichlet(\alpha_1 + \beta, \alpha_2 + \beta, \alpha_3 + \beta, \alpha_4 + \beta),$$

where $\alpha_1, \ldots, \alpha_4$ are the number of markers in each component.

New genetic variances of the mixture components $\sigma_k^2$ are computed

$$\sigma_k^2 \sim \begin{cases} \pi_1 \times 0 \times \sigma_g^2 \\ \pi_2 \times 10^{-4} \times \sigma_g^2 \\ \pi_3 \times 10^{-3} \times \sigma_g^2 \\ \pi_4 \times 10^{-2} \times \sigma_g^2 \end{cases}.$$

It becomes clear from the last step that the additive genetic variances of the mixture components are not estimated in this implementation of BayesR. The default implementation makes use of arbitrary values of $0, 10^{-4}, 10^{-3}, 10^{-2}$ as scaling to get the additive genetic variances for markers assigned to the mixture components respectively. The total additive genetic variance for a mixture component $k$, $\sigma_k^2$ is then computed by scaling the total additive $\sigma_g^2$ with the total number of markers assigned to that component $\pi_k$ and the arbitrarily assigned scaling value for the component.

To investigate the robustness of BayesR I tested a variant of the model in a simulation study in which I used the true scaling values obtained from variances simulated for SNPs in each mixture component instead of the arbitrary values. These scaling values are $0, 4 \times 10^{-5}, 6 \times 10^{-4}, 0.025$ for low heritability traits, $0, 10^{-4}, 6 \times 10^{-4}, 0.01$ for moderate heritability traits and $0, 1.25 \times 10^{-4}, 5 \times 10^{-4}, 6.25 \times 10^{-3}$ for high heritability traits. These true values were used as scaling values for the additive genetic variances for SNPs assigned to the mixture components respectively. Also, I ran a model that changed the starting values for the Dirichlet distribution for the proportions of SNP in each mixture component from the default values of $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3, \alpha_4) = 1$ to $\boldsymbol{\alpha} = (1,1,1,10)$.

The order in which the markers are resampled is randomly permuted in each iteration to improve the mixing. The analysis was done using a Markov Chain Monte Carlo (MCMC) chain length of 50,000 with the first 20,000 chains being discarded as burn-in. Every 10$^{th}$ iteration after burn-in was kept to create visual plots of the sampling.

### 3.2.3 Simulation of phenotypes

In order to evaluate the model performance and also the accuracy of the model parameters estimated by BayesR, I performed a simulation study. Phenotypes were simulated using available genotypic information of the 2,896 individuals with urine phenotype measurements in the Generation Scotland: Scottish Family Health Study (GS: SFHS) (Smith et al., 2012). A quality control (QC) that removed SNPs with MAF $< 0.01$ and SNPs that were out of Hardy-Weinberg equilibrium at p-value $< 0.00001$ was performed. A total of 555,091 SNPs remained after QC.

Three phenotypes with high, moderate and low heritability of 80%, 50% and 10% respectively were simulated. For simulating the SNP effects, I assumed they came from four normal distributions with zero means and variances that were equal to their contribution to the total additive genetic variance. The first distribution contributed nothing to the additive genetic effect. For the second distribution, 5,000 SNPs were randomly sampled to form the background polygenic effect, then another 500 SNPs were randomly sampled to have moderate SNP effects in the third distribution. A further 20 SNPs were sampled randomly to have large SNP effects in the final distribution. The small effect sizes were drawn from a $N(0,0.5)$, $N(0,0.25)$, and $N(0,0.02)$ respectively for the high, moderate and low heritability phenotypes. The moderate effect sizes were drawn from a $N(0,0.2)$, $N(0,0.15)$ and $N(0,0.03)$ for the high, moderate and low heritability phenotypes. The large effect sizes were drawn from a $N(0,0.1)$, $N(0,0.1)$ and $N(0,0.05)$ for the high, moderate and low heritability phenotypes respectively.

The effect $g_j$ of a SNP $j$ from a distribution component $k$ under a heritability class was calculated from the additive genetic variance $\sigma_k^2$ contribution from the respective distribution as follows:

$$\sigma_{kj}^2 = 2p_j(1-p_j)g_j^2, \tag{3.9a}$$

$$g_j = \sqrt{\frac{\sigma_{kj}^2}{2p_j(1-p_j)}}, \tag{3.9b}$$

where $p_j$ is the frequency of the effect allele of the SNP $j$.

A random environmental variance was simulated for each individual for the phenotypes. The environmental variances were derived from $N(0, \sigma_e^2)$ where $\sigma_e^2$ is 0.2, 0.5 and 0.9 for the high, moderate and low heritability traits. The final simulated phenotype for an individual $i$ was then calculated as follows

$$y_i = \sum_{j=1}^{5000} x_{ij}g_j + \sum_{j=1}^{500} x_{ij}g_j + \sum_{j=1}^{20} x_{ij}g_j + e_i, \tag{3.10}$$

where $x_{ij}$ is the number of copies of the effect allele of SNP $j$ and $g_j$ is the effect of SNP $j$. Thus, the final phenotype for each heritability class was the sum of the additive genetic variance from the three distributions and the residual variance, with an expected mean of zero and a variance of 1. Fifty replicates were analysed for each of the three heritability classes with a different set of SNPs sampled for each replicate.

### 3.2.4 GBLUP

The simulated phenotypes were also analysed using a GBLUP model. The model below was fitted to the data

$$y = 1\mu + Xu + e, \tag{3.11}$$

where $\boldsymbol{y}$ is a vector of $n$ observations for the phenotype, $\boldsymbol{\mu}$ is the overall mean effect which is fixed; considering a vector of $m$ genome-wide SNPs, $\boldsymbol{X}$ is an $n \times m$ design matrix that assigns phenotypes to marker effects; $\boldsymbol{u}$ is a vector of $m$ marker effects assumed to be multivariate normal, $u \sim MVN\left(0, \boldsymbol{G}\sigma_g^2\right)$; $\boldsymbol{e}$ is a vector of $n$ residual effects, $\boldsymbol{e} \sim MVN(0, \boldsymbol{I}\sigma_e^2)$ with $\boldsymbol{I}$ being an $n \times n$ identity matrix. $\boldsymbol{G}$ is the GRM, calculated from genome-wide SNPs that is scaled to be analogous to the pedigree-based numerator relationship matrix used in BLUP (VanRaden, 2008).

$$G = \frac{ZDZ'}{m} \tag{3.12}$$

The elements in column $j$ $\left(z_{ij}\right)$ of $\boldsymbol{Z}$ can take the values $\left(0 - 2p_j\right), \left(1 - 2p_j\right), and\ \left(2 - 2p_j\right)$ for the homozygote genotype with the reference allele, the heterozygote genotype and the homozygote genotype with the effect allele; $p_j$ is the reference allele frequency at locus $j$; $\boldsymbol{D}$ is a diagonal matrix that scales $\boldsymbol{G}$**.** The scale used in this model is the locus specific scale described by VanRaden (2008)

$$d_{jj} = \frac{1}{2p_j\left(1 - p_j\right)} \tag{3.13}$$

### 3.2.5 Prediction analysis to assess model performance

One of the benefits of a GWA analysis model is its ability to predict breeding values or risks when there are no phenotype data available. Prediction usually informs economically important decisions, especially in animal breeding which makes it imperative that some form of quality assessment of the model is done before its implementation.

To assess how good a model is for prediction of new phenotypes, based on the data available, it is common to perform a $k$-fold cross-validation, where the data is split into $k$ subsets. The analysis then uses each of the $k$ subsets as a validation set to predict phenotypes using information from the other $k-1$ subsets of the data.

Two parameters are measured in the validation set to assess predictive performance, and these are accuracy and bias. Accuracy is the correlation of the estimated breeding values (EBVs) with the true breeding values (TBVs). I, therefore, calculated accuracy as the correlation between the simulated breeding values and the EBVs. I selected 30 of the simulated phenotypes, 10 from each heritability class for the prediction analysis. For each replicate, phenotypes were randomly split into five subsets to be used in a 5-fold cross validation analysis to evaluate the predictive performance of the BayesR and GBLUP models.

## 3.3 **Results**

One major drawback of the standard GWA analysis is that each marker is analysed one at a time ignoring the background effect of all other markers. This has been pointed out as problematic (de los Campos et al., 2010; Hoggart et al., 2008) and the way around this problem is the use of methods that analyse all markers at the same time.

I tested two models that analysed markers all at once using phenotypes I had simulated under three different heritability classes. The first model is a Bayesian mixture model that assumes a mixture of four normal distributions with zero means as the prior distributions for sampling the effect of a particular SNP. These

distributions are such that they contribute increasing proportions of the total additive genetic variance. The model samples the estimates of four parameters from their conditional posterior distributions using a Gibbs sampling scheme. These parameters are the additive genetic variance, the residual variance, the overall mean and the number of SNPs that go into each of the mixture distributions.

### 3.3.1 BayesR parameter trace

A Gibbs sampler uses a Markov Chain to make Monte Carlo approximations of the posterior distribution of parameters by drawing a large number of samples from it. To visualise how the parameters are sampled from the posterior distribution by the MCMC algorithm I produce trace plots of how the chain treads through the parameter space. These plots provide evidence of how the Markov chain converges to the posterior distribution. A good MCMC algorithm allows the parameters to have sufficient state changes along the chains. This is indicative of proper mixing, which means the algorithm jumps around well in the parameter space.

I ran an MCMC chain of 50,000 cycles for each BayesR model. The first 20,000 cycles of the chain were discarded as burn-in. Plots of the posterior samples of the additive genetic variance, the error variance and the number of SNPs sampled

for each mixture component were generated to visualise the sampling.

The plots in Figure 3.1 to Figure 3.4 were generated by drawing from every $10^{th}$ sample after burn-in to give a total of 3,000 posterior samples. The plots in Figure 3.1 show that there are no mixing problems for the sampling of both the genetic and error variance for any of the heritability traits under any of the BayesR models. In the

low heritability traits, the trace for the posterior number of SNPs (Figure 3.2) shows that for the $k1\ and\ k2$ components, the algorithm stays trapped in certain parameter states for longer periods at a time before jumping to other states. The trace of the two component distributions mirrors each other.

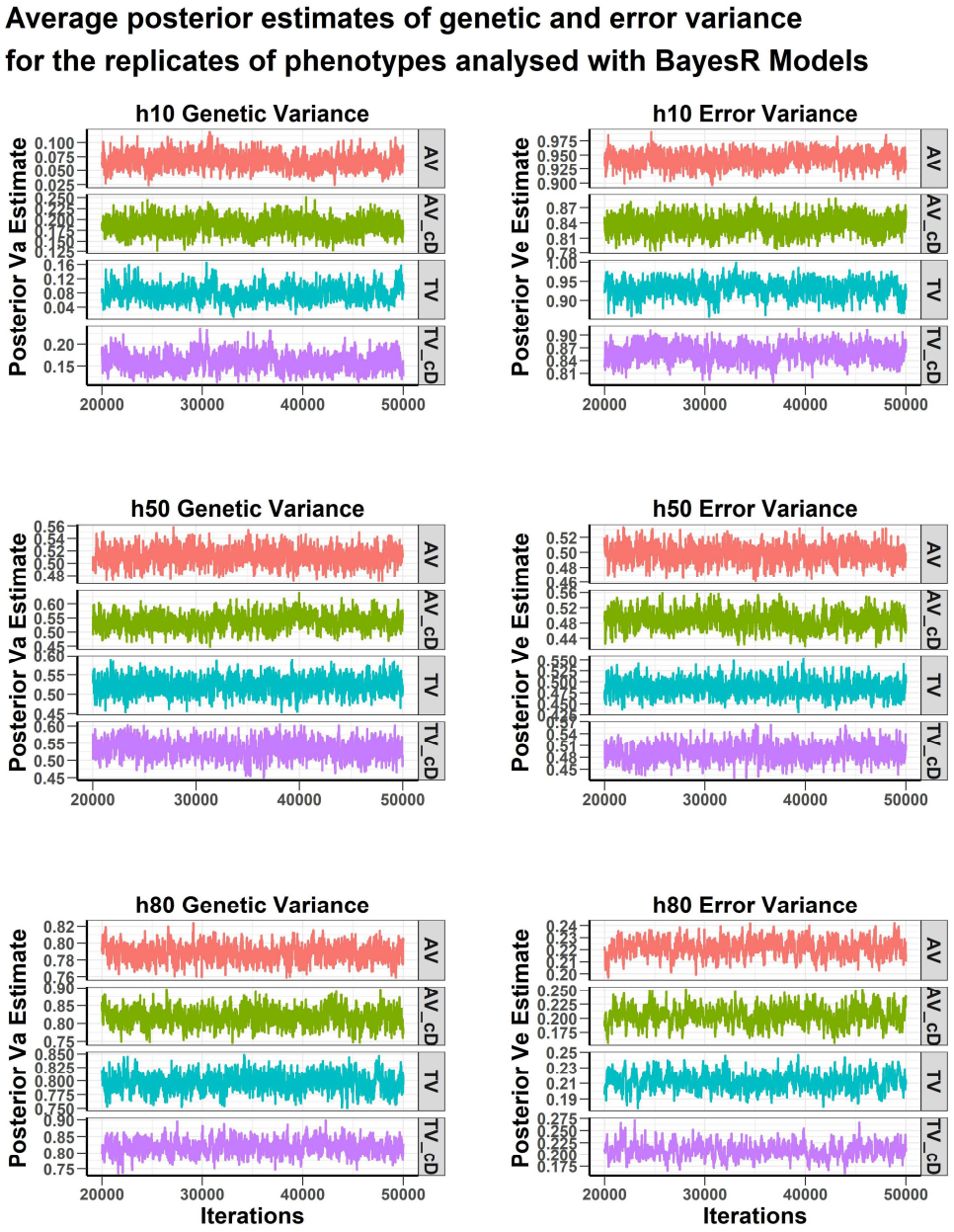



Figure 3.1. Trace-plots of the posterior estimates of the genetic and error variance for each heritability trait for the 4 BayesR models. The MCMC chain was run for 50,000 cycles with the first 20,000 cycles discarded as burn-in. The trace-plots are therefore based on 3000 sampled posterior parameters estimated drawn at every 10th cycle after burn-in. The plots show that the MCMC algorithm had no mixing problem when sampling the two parameters.

This means the two parameters are correlated and this is what potentially leads to the bad mixing. Perhaps this correlation is being driven by the fact the SNPs in these distributions contribute nothing and very little (about $10^{-4}$ on average per SNP) to the additive genetic variance respectively.

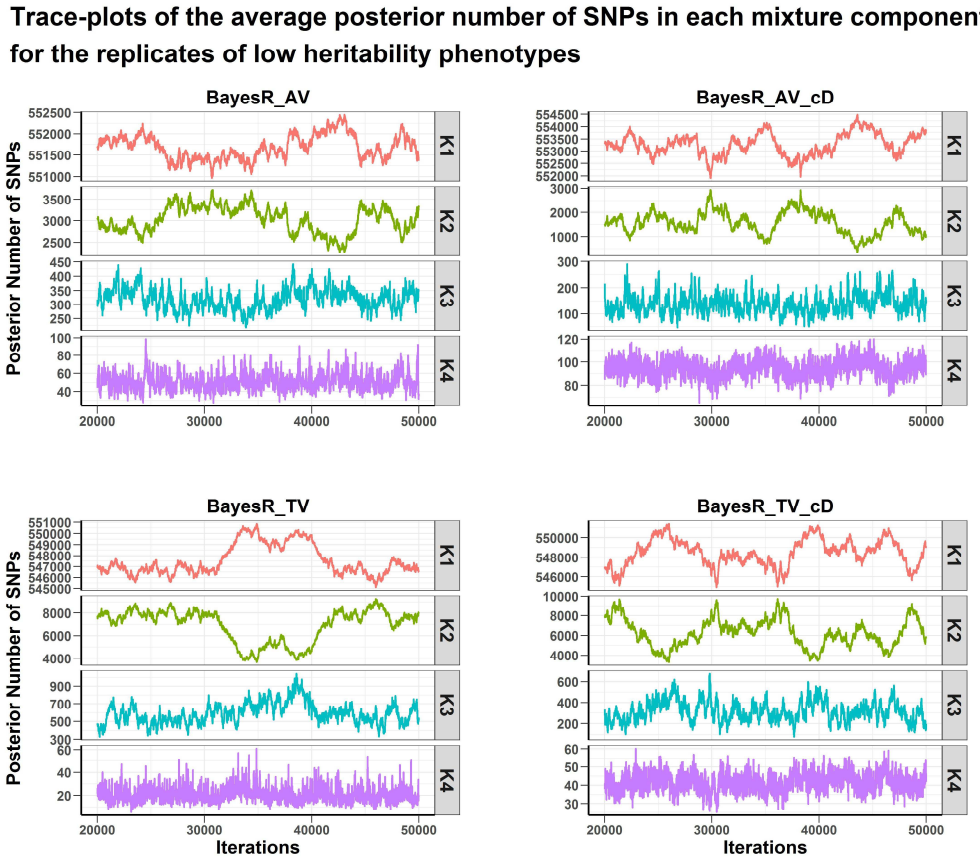



Figure 3.2. Trace-plots of the posterior estimates of the number of SNPs in each component of the mixture distributions for the low heritability traits for the 4 BayesR models. The MCMC chain was run for 50,000 cycles with the first 20,000 cycles discarded as burn-in. The trace-plots are therefore based on 3000 sampled posterior parameters estimated drawn at every $10^{th}$ cycle after burn-in. The plots for the $k3\ and\ k4$ components show that the MCMC algorithm had no mixing problems.

The mixture model is a combination of the four components and thus without sufficient evidence to distinguish the $k1\ and\ k2$, the algorithm wouldn't be able to correctly classify SNPs in those two distributions. So, at one realisation it will put a particular group of SNPs in one distribution and at another realisation it will place

those SNPs in the other distributions. The $k3\ and\ k4$ components however, show very good mixing (Figure 3.2).

The trace of the moderate (Figure 3.3) and high heritability traits (Figure 3.4) show a correlation between the parameter spaces for the first 3 components of the mixture distribution. The last component, however, shows better mixing.

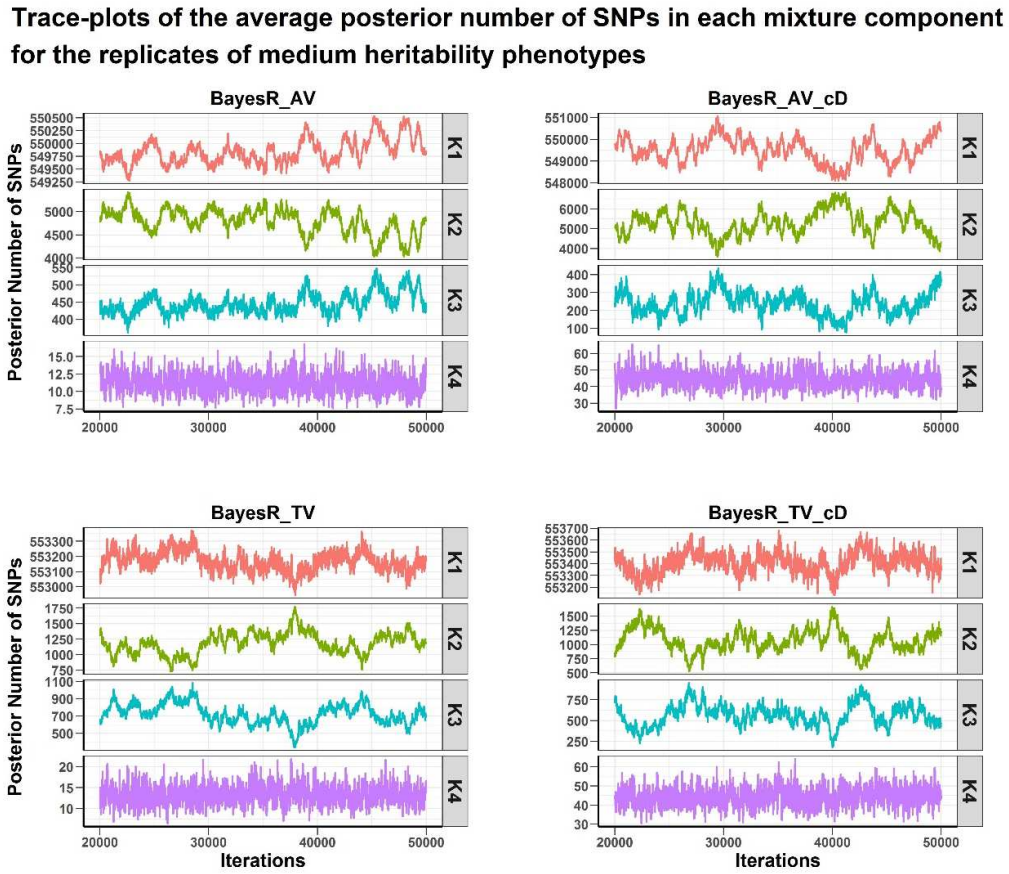



Figure 3.3. Trace-plots of the posterior estimates of the number of SNPs in each component of the mixture distributions for the medium heritability traits for the 4 BayesR models. The MCMC chain was run for 50,000 cycles with the first 20,000 cycles discarded as burn-in. The trace-plots are therefore based on 3000 sampled posterior parameters estimated drawn at every 10$^{th}$ cycle after burn-in. The plots $k4$ components show that the MCMC algorithm had no mixing problem. The $k1, k2\ and\ k3$ components show bad mixing for the sampling.

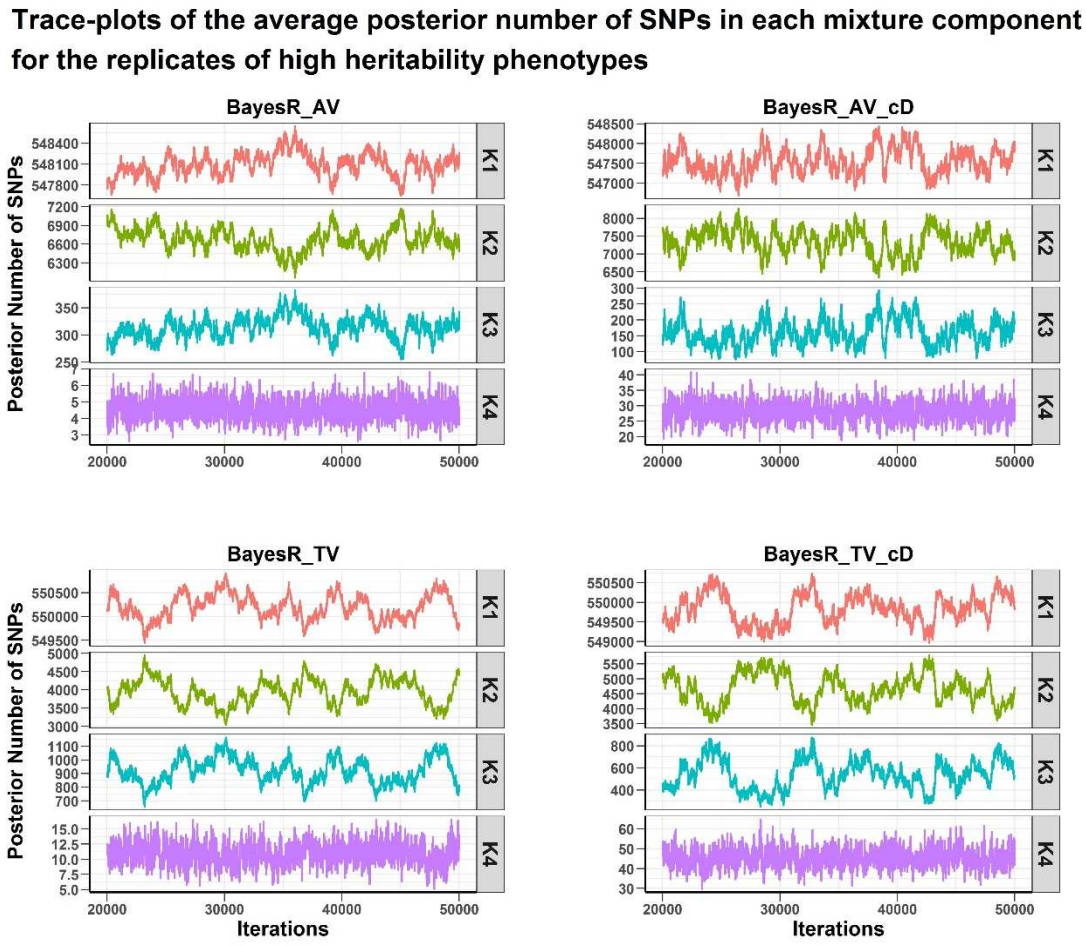



Figure 3.4. Trace-plots of the posterior estimates of the number of SNPs in each component of the mixture distributions for the high heritability traits for the 4 BayesR models. The MCMC chain was run for 50,000 cycles with the first 20,000 cycles discarded as burn-in. The trace-plots are therefore based on 3000 sampled posterior parameters estimated drawn at every 10$^{th}$ cycle after burn-in. The plots $k4$ components show that the MCMC algorithm had no mixing problem. The $k1, k2\ and\ k3$ components show bad mixing for the sampling.

### 3.3.2 Estimates of model parameters by the two models

Estimates of the additive genetic and residual variances from this model were compared to those obtained from a GBLUP model. These results are summarised in Figure 3.5 and Table 3.1.

I simulated 50 replicates of three different phenotypes with differing heritability values from low to high having 10%, 50% and 80% heritability, represented by h10, h50 and h80 in the results, respectively.

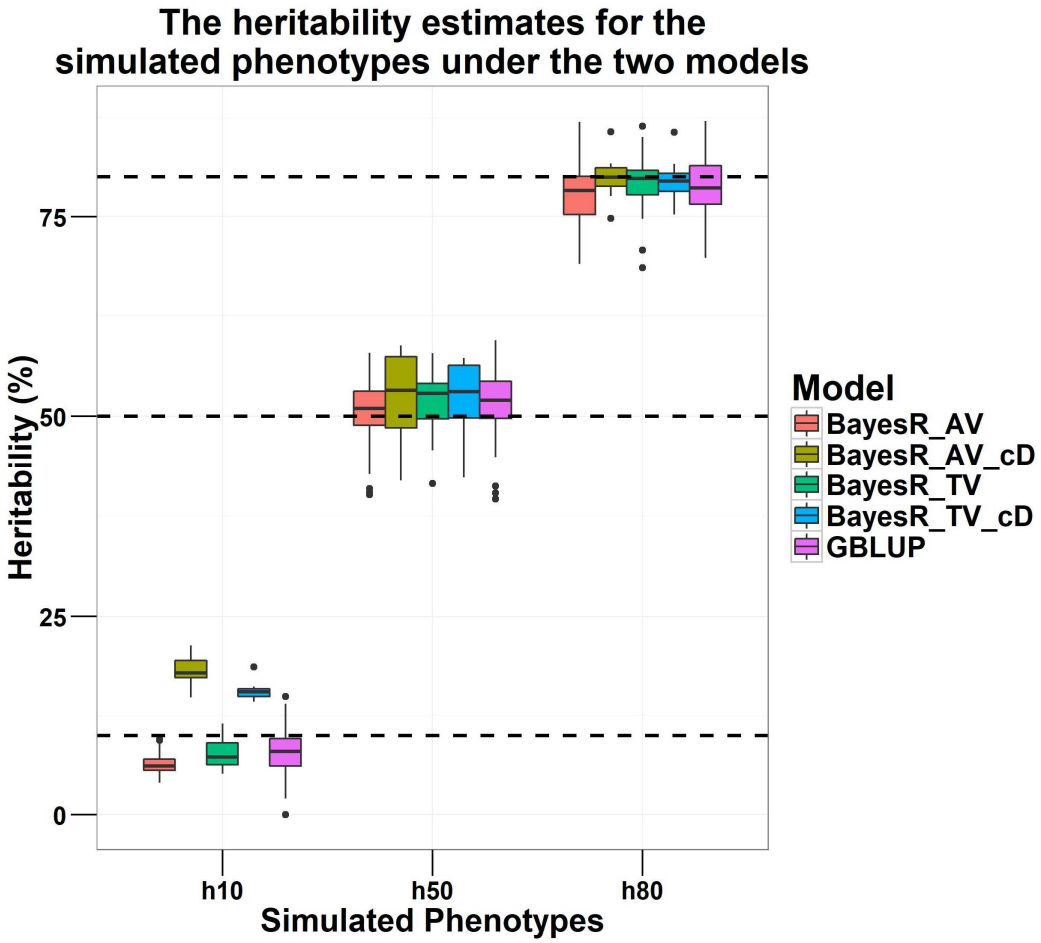



Figure 3.5. Comparison of the heritability estimates of the simulated phenotypes obtained from the two models. BayesR with a default variance value for SNPs in the mixture components (BayesR_AV), BayesR with true values of SNP variance simulated for the three heritability traits (BayesR_TV), and BayesR under the first two instances but with starting values for the Dirichlet priors changed (BayesR_AV_cD and BayesR_TV_cD). The dashed lines are simulated heritability values for the different traits. The heritability estimates for the BayesR models with Dirichlet priors changed were inflated for the low heritability traits.

The additive genetic effects for these traits were simulated using a random sample of SNPs across the genome totalling 5,520. These SNPs were split across three normal distributions with zero means and varying proportions of contribution to the total additive genetic variance of each simulated phenotype. The remaining SNPs were assumed to have no effect on the traits and thus contribute nothing to the total additive genetic variance. These zero effect SNPs constituted another distribution for

each of the simulated traits to give a total of four normal distributions as conditional priors for sampling the effects of SNPs. This simulation was in line with the basic assumption of BayesR.

The BayesR model was implemented in four different ways. First was a BayesR model with arbitrarily set values for scaling the variance for SNPs in the mixture components, represented in the results as BayesR_AV. The second was a BayesR model which utilised true scaling values for the SNP variance simulated under the three heritability classes (BayesR_TV), and the last two were BayesR models with similar variance scaling values to the first two instances but with starting values for the number of SNPs that goes into the mixture components changed i.e. the Dirichlet priors were changed (BayesR_AV_cD and BayesR_TV_cD).

The results reported in Table 3.1 and in Figure 3.5 show that the BayesR_AV and BayesR_TV models underestimated the heritability of h10 traits by about 35% and 22% respectively, on average. However, the heritability of the h10 traits analysed with BayesR models with Dirichlet priors changed were overestimated on average. The heritability for the h50 and h80 traits were estimated fairly accurately by all models, although some models slightly underestimated or overestimated them by between 1 – 5%.

A total of 5,000, 500 and 20 SNPs were simulated to have an effect in the $k2, k3\ and\ k4$ components of the mixture distributions respectively. It can be seen from Table 3.1 that analysing the low heritability traits (h10) with the default implementation of BayesR (BayesR_AV) underestimated the number of SNPs that

went into the $k2\ and\ k3$ components by almost half whilst overestimating the number that went into the $k4$ component by more than double on average.

Table 3.1. The posterior estimates of model parameters by the BayesR models for the three heritability classes. The model name, simulated trait name, heritability (calculated using the posterior estimates of the genetic and residual variances), estimated total number of SNPs, estimated number of SNPs for the second mixture component, estimated number of SNPs for the third component of the mixture distributions and the estimated number of SNPs for the fourth component of the mixture distributions. The values reported are the mean and standard deviation in the parenthesis of each estimate across the replicates of each heritability class. The $k2, k3\ and\ k4$ components are simulated to each contribute to the total additive genetic variance.

| Model | Trait | % heritability | nSNPs | $(k2)$ 5000 SNPs | $(k3)$ 500 SNPs | $(k4)$ 20 SNPs |
|---|---|---|---|---|---|---|
| BayesR_AV | h10 | 6.51 (1.37) | 3402 (749) | 3031 (796) | 319 (46) | 50 (7) |
| | h50 | 50.61 (4.25) | 5260 (1228) | 4804 (1348) | 444 (125) | 11 (5) |
| | h80 | 78.3 (3.96) | 7019 (1026) | 6704 (1138) | 311 (113) | 4 (2) |
| BayesR_AV_cD | h10 | 18.2 (1.86) | 1810 (512) | 1580 (511) | 137 (17) | 94 (5) |
| | h50 | 52.4 (5.69) | 5590 (845) | 5300 (899) | 239 (70) | 44 (7) |
| | h80 | 80 (2.98) | 7540 (610) | 7350 (664) | 161 (53) | 28 (3) |
| BayesR_TV | h10 | 7.73 (1.69) | 7540 (2645) | 6930 (2762) | 591 (136) | 21 (3) |
| | h50 | 51.8 (4.25) | 1920 (105) | 1180 (203) | 726 (148) | 13 (5) |
| | h80 | 79 (4.26) | 4890 (996) | 3950 (1304) | 923 (316) | 11 (7) |
| BayesR_TV_cD | h10 | 15.7 (1.42) | 6640 (1587) | 6280 (1605) | 316 (63) | 41 (3) |
| | h50 | 52 (4.99) | 1680 (146) | 1070 (195) | 566 (92) | 44 (7) |
| | h80 | 79.7 (2.86) | 5250 (385) | 4670 (475) | 540 (102) | 45 (6) |

When the Dirichlet priors were changed, a similar trajectory was observed with the number that went into the $k4$ component more than quadrupled on average for the low heritability traits. In the moderate heritability traits (h50), these two implementations of the BayesR model estimated the number of SNPs that got sampled into the $k2\ and\ k3$ components fairly well on average but underestimated and overestimated the numbers that went into the $k4$ component for the

BayesR_AV and BayesR_AV_cD models by half respectively. Both models overestimate the number of SNPs that are sampled into the $k2$ components and underestimate the $k3$ numbers. The $k4$ numbers are grossly underestimated for h80 in the BayesR_AV model and overestimated by half in the BayesR_AV_cD.

Further implementations of the BayesR model involved the use of the true scaling values obtained from the additive genetic variance simulated for SNPs in the various components of the mixture. Table 3.1 shows that this implementation of the model may not be very different from the default implementation. The only thing that stands out is the gross underestimation of the number of SNPs that are placed in the $k2$ component of the mixture distributions for the moderate heritability traits.

Two things become obvious from these results. The first is that using arbitrary or true scaling values of genetic variance of SNPs for the mixture components in BayesR, on the average, underestimates the additive genetic variance of low heritability traits (Figure 3.5) except when the values for the Dirichlet priors are changed in which case it overestimates it. Second is that there is not much difference in the estimation of the number of SNPs with effects under the different models of BayesR. The importance of this is that there is not much benefit to be gleaned when the BayesR model is fitted with true prior values as opposed to arbitrary values which most likely will be the case in analysing real phenotypes.

### 3.3.3 Selection of effect SNPs by the two models

The heritability and the number of effect SNPs estimated by the models are remarkably not far off from the true numbers simulated on average. But more importantly, are the models finding the same effect SNPs that were simulated and

are they being placed in the right component class? To answer this question, nine simulated traits for each heritability class were chosen to investigate if the BayesR models pick simulated effect SNPs and if they place SNPs into the correct mixture distribution. The nine simulated traits selected for each heritability class were those with the lowest, middle and maximum values for estimated heritability, the total number of effect SNPs and number of SNPs in the $k4$ distribution in the h50 heritability traits. In total, 27 simulated traits were selected across the three heritability classes.

For each of the chosen simulations, I ranked all SNPs based on the estimated effect sizes and picked the top-ranked SNPs based on the total number of SNPs reported to have an effect by the models. I used the absolute values of the effect sizes for ranking SNPs without regard to their sign. I compared the RS-IDs of the top SNPs to those in the list of SNPs simulated to have an effect. The common SNPs between the two lists were counted for all the 5520 SNPs simulated to have an effect and another count that involved only the 20 SNPs simulated to have large effects. The same analysis was done for the GBLUP model and the results are shown in Figure 3.6.

The total number of overlaps between estimated and simulated SNPs with effect was on average very low across models and traits. The numbers improved when moving from low heritability traits to high heritability traits. BayesR on average performed better than GBLUP in the high heritability traits. The BayesR models on average were able to recover almost all the 20 SNPs with large effects in the medium and high heritability traits. GBLUP only managed to recover half of these large effect

SNPs on average in the medium and high heritability traits. The benefit of using the true SNP variances in a BayesR model was observed only in low heritability traits, whereas in the medium and high heritability traits it drove the numbers of true effect SNPs captured down. So again, there is no real benefit in using the true values of the SNP variances.

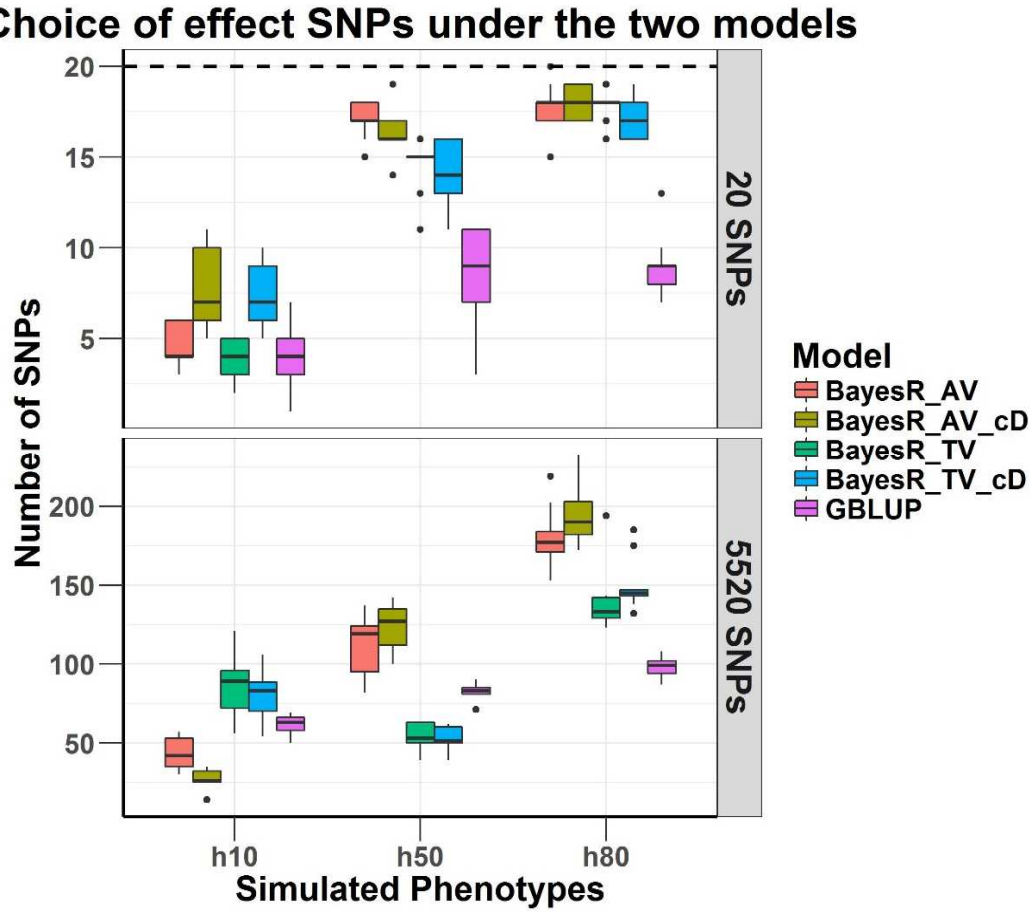



Figure 3.6 Models choice of SNPs with effect over nine traits. The upper panel gives information on the top 20 SNPs and the lower panel is the total number of SNPs simulated to have an effect. The x-axis is the three different heritability traits. The BayesR models perform better than the GBLUP model in the moderate and high heritability traits. Both models perform similarly in the low heritability traits. In total, only a handful of the SNPs simulated to have an effect were picked up by the models. But on average the BayesR models were able to pick up nearly all the 20 SNPs simulated to have the largest effects in the moderate and high heritability traits

### 3.3.4 Genetic architecture of traits

One of the selling points of BayesR is that it lends itself to the study of the genetic architecture of traits. The whole genome architecture of the simulated traits was investigated to assess the performance of BayesR in dissecting the genetic architecture of traits. Where genetic architecture refers to the number of loci that affects a trait and the effect sizes of these loci. The results for the genetic architecture of the simulated traits are given in Figure 3.7 as proportions of the additive genetic variance explained by SNPs with effect. These are averaged over the 50 simulations for each heritability class.

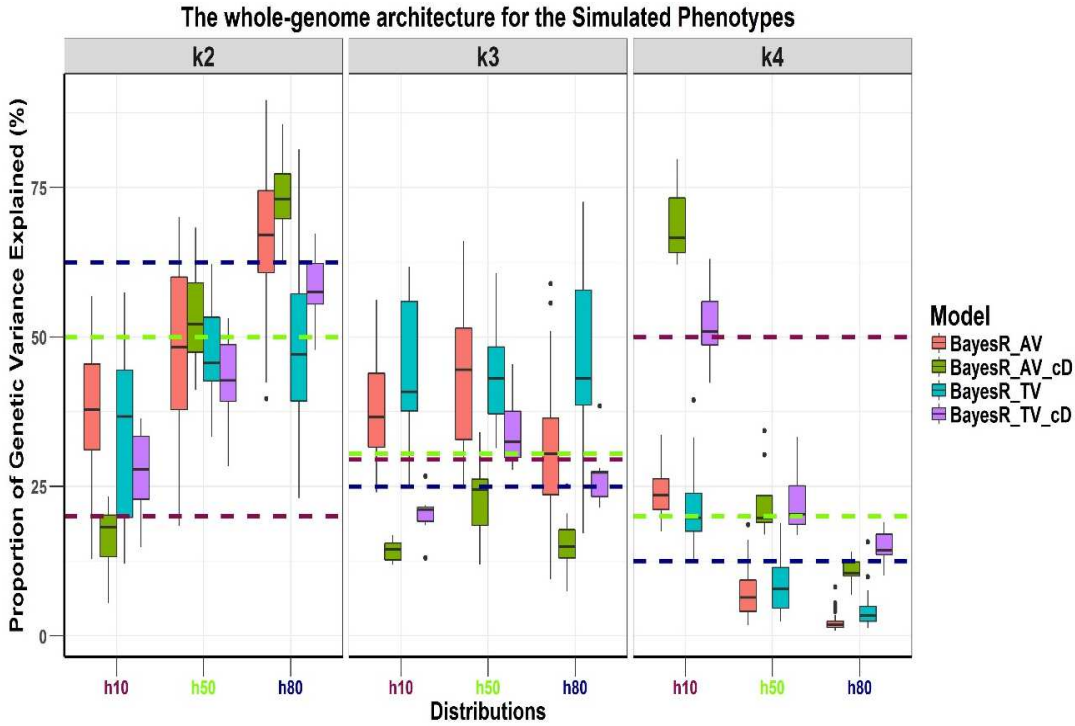



Figure 3.7. The whole genome architecture of the different heritability traits estimated using the different BayesR models. The proportion of additive genetic variance explained by SNPs with effect. The plot facets represent the three distributions describing the SNP effects. The dashed lines are the proportions of genetic variance simulated for the three traits; they are colour coded to correspond with the respective heritability traits. The medium and high heritability traits are estimated to have polygenic architecture. The two BayesR models with the Dirichlet prior changed on a whole gave better estimates of the proportion of variance explained than the other two models.

The h10 trait was simulated such that SNPs in the $k2, k3\ and\ k4$ components of the mixture distributions accounted for 20%, 30% and 50% of the additive genetic variance respectively. BayesR_AV estimated these to be 37.8%, 37.9% and 24.2% on average respectively. When the Dirichlet prior was changed for the $k4$ component, the estimates were 16.82%, 14.30% and 68.46% on average. In the model that used the true SNP variances, changing the Dirichlet prior improved the estimates of the genetic variances in the mixture components.

The h50 trait was simulated to have a polygenic architecture (bigger contribution to the total additive genetic variance from the low variance SNP distribution) and consequently, the $k2, k3\ and\ k4$ components of the mixture distributions explained 50%, 30% and 20% of the additive genetic variance respectively.

The BayesR models estimated these proportions fairly well. The models with the changed Dirichlet priors outperformed the models that used the default uninformative priors. The BayesR models were able to estimate the proportions of the $k2$ distribution close enough but the $k3$ proportions were slightly overestimated whereas the $k4$ proportions were underestimated in the models with default uninformative Dirichlet priors.

The h80 trait also followed a polygenic architecture and the $k2, k3\ and\ k4$ components of the mixture distributions were simulated to account for 62.5%, 25% and 12.5% of the additive genetic variance respectively. BayesR models could accurately determine the genetic architecture for the high heritability traits. Similar

results were observed in the high heritability traits as in the moderate heritability traits. The $k4$ proportions of the variance were underestimated in the models with uninformative Dirichlet priors.

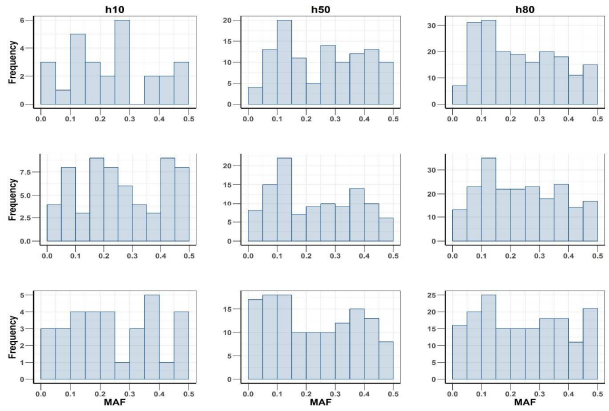
### 3.3.5 Investigating the relationship between MAF, effect sizes and posterior inclusion probability

I investigated the MAF spectrum for the SNPs with estimated effect, the relationship between the MAF, the effect sizes and the posterior inclusion probability using the BayesR_AV model for the nine selected traits as described in the section above. The SNPs with effects are well distributed across the MAF spectrum for all the heritability traits (Figure 3.8i – iii).

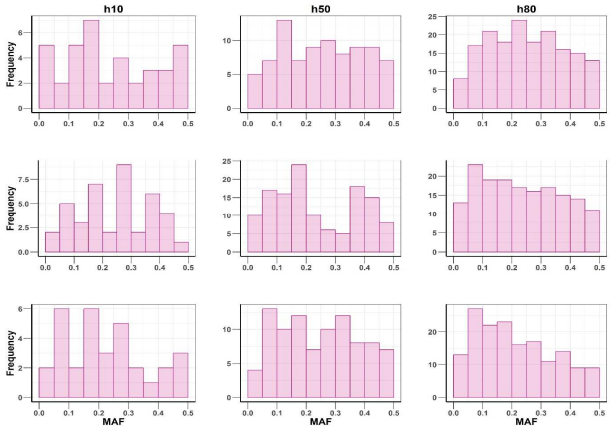
There seems to be no apparent relationship between the MAF of SNPs and the SNP effects, and the posterior probability of a SNP being included in $k2, k3\ or\ k4$ components of the mixture distributions, Figure 3.9 – Figure 3.10.

I further investigated whether the SNPs with effect detected in the BayesR model are located within regions where I simulated SNPs to have an effect. I employed a rolling non-overlapping window approach along defined regions of 1000 SNPs spanning the whole genome. So, for each region of 1000 SNPs, I calculated the sum of the absolute values of their simulated effects. I compared these sums to the sums of the absolute values of estimated SNP effects obtained using the BayesR model. The results of this analysis are shown in Figure 3.11 for the three different heritability traits. The sums of the estimated SNP effects are lower in magnitude and are spread uniformly across the genome as compared to the sums of simulated effects.

i.




ii.




iii.

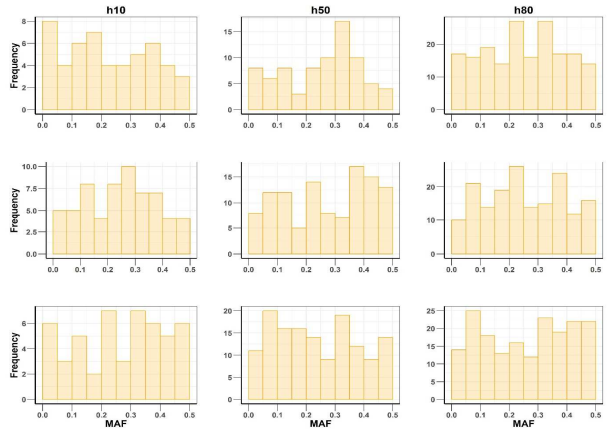
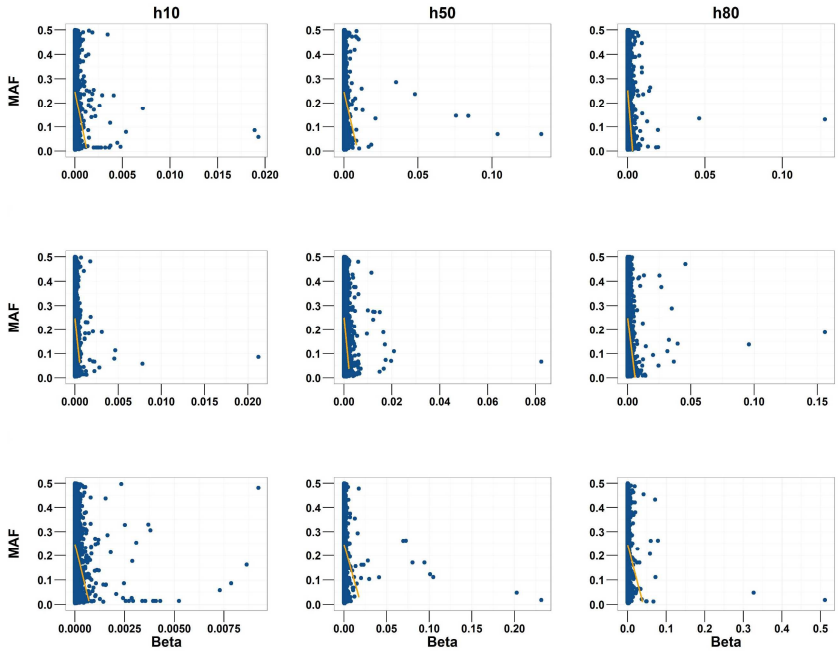



Figure 3.8. The minor allele frequency spectrum of SNPs sampled to have an effect on traits. The panels represent a total of nine traits chosen for each heritability level. The selection criteria used is such that the top panels (i) are chosen based on the estimated heritability, the middle panels (ii) are based on the number of SNPs sampled for the $k4$ distribution and the lower panels (iii) are based on the total number of SNPs sampled for the traits. The plots show that there is no apparent relationship between MAF and estimated SNPs with effects. BayesR does not discriminate on what SNP to sample to have an effect based on its MAF but samples across the whole MAF spectrum.

i.




ii.

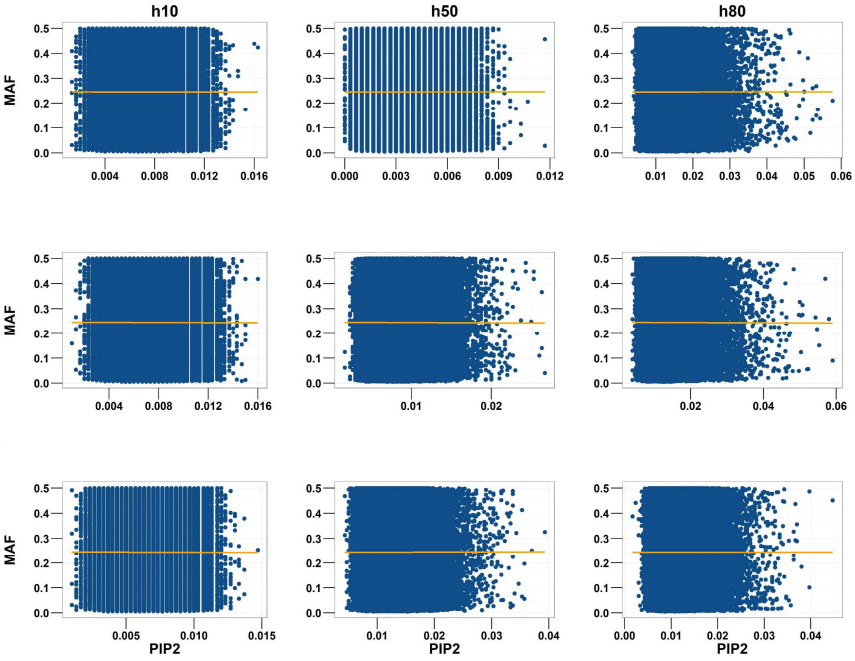



Figure 3.9. The relationship between minor allele frequency (MAF) of genome-wide SNPs and the estimated SNP effects (i) and posterior inclusion probability for the $k2$ distributions (PIP2) (ii). Three traits were selected for each heritability. The plots show that there is no relationship between MAF of SNPs and the SNP effects or the PIP2.

i.

ii.

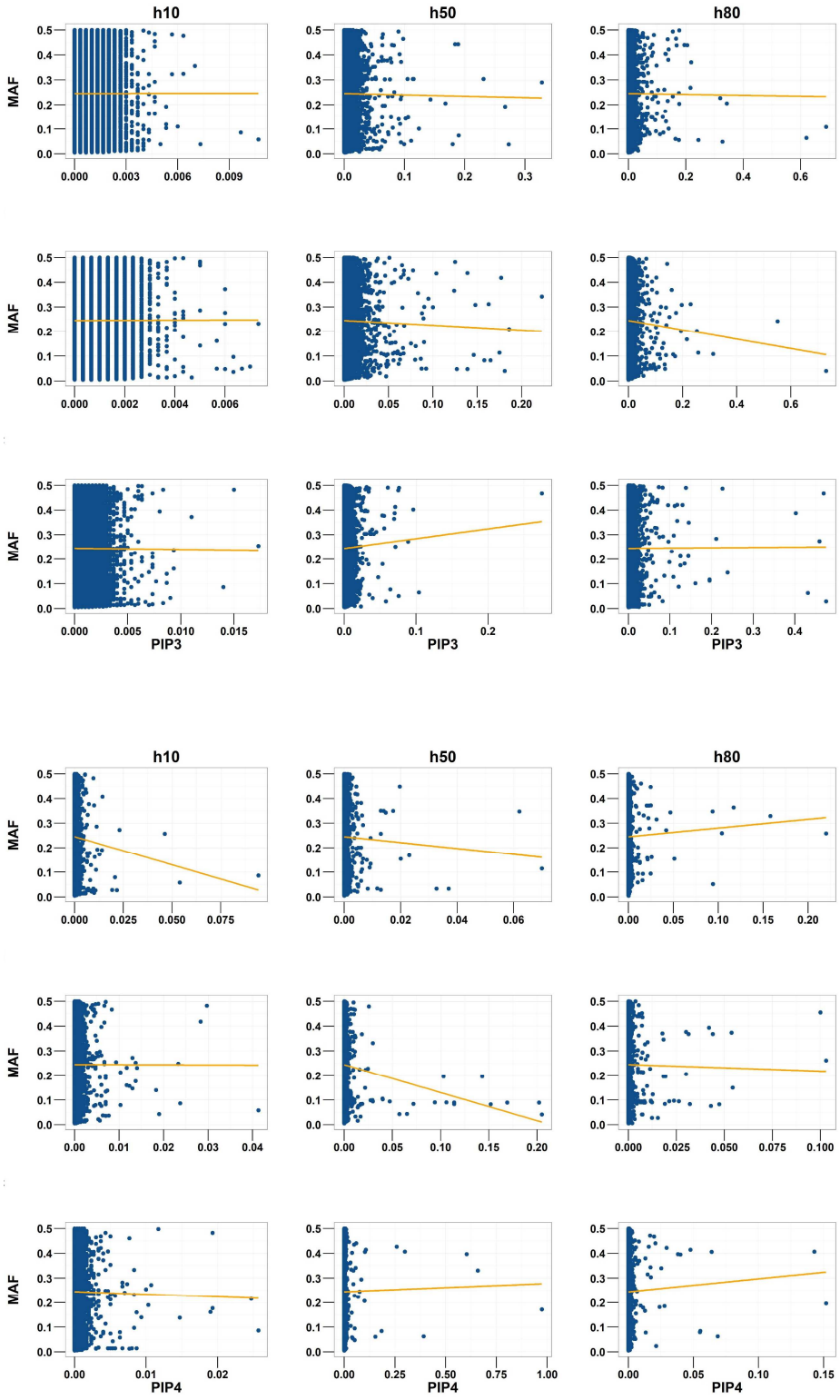



Figure 3.10. The relationship between minor allele frequency (MAF) of genome-wide SNPs and the posterior inclusion probability for the $k3$ $and$ $k4$ distributions (PIP3 and PIP4). Three traits were selected for each heritability. The plots show that there is no relationship between MAF of SNPs and the PIP3 or the PIP4.

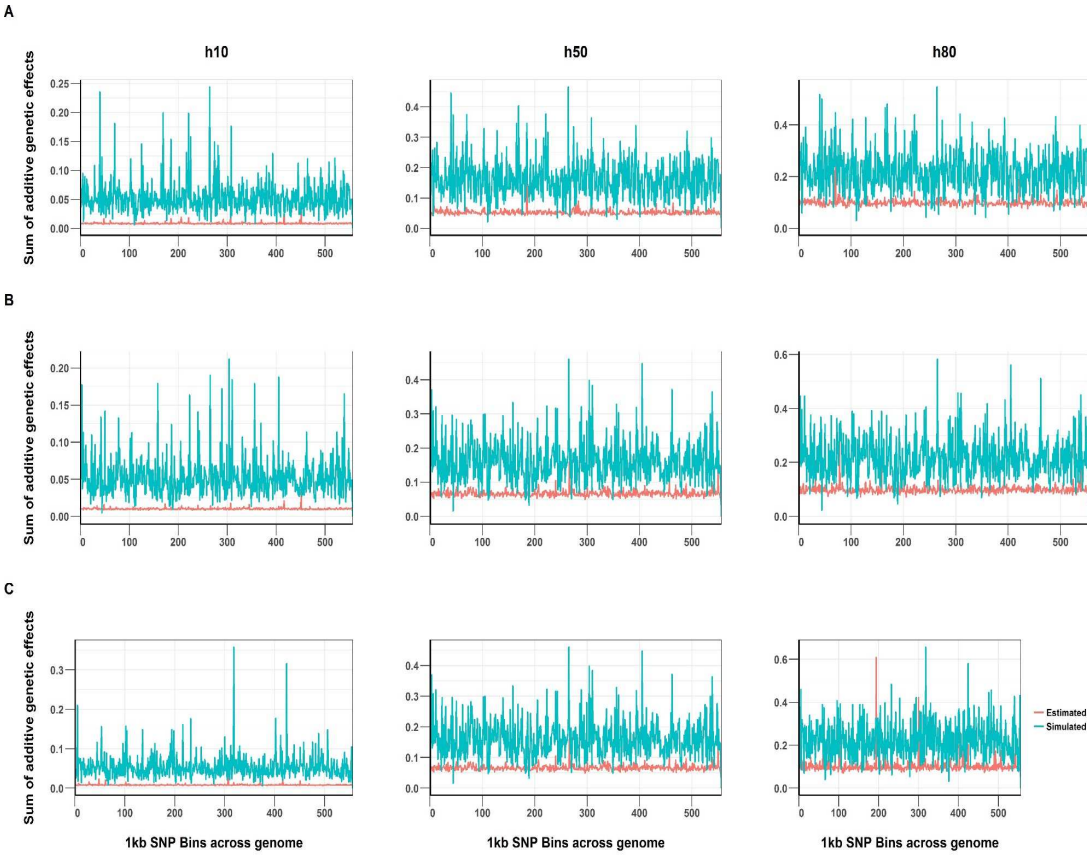



Figure 3.11. Comparison between the simulated SNP effects and the BayesR estimates of SNP effects. The plots are sum of effects in 1000 SNP windows across the genome. The estimated SNP effects shown in red are lower on average across the regions than the simulated SNP effects.

### 3.3.6 Models prediction

To evaluate the accuracy of phenotype prediction from genotype data using the BayesR and GBLUP models, I performed a cross-validation analysis using 30 of the simulated sets of phenotypes, 10 from each of the three heritability classes. The data within each replicate of the simulation were randomly split into five subsets to be used in 5-fold cross-validation analysis. The default BayesR model was used in this prediction analysis. For both models, the accuracy of prediction was based on the estimated effects of all markers. The accuracy of the prediction reported in Table 3.2 is the correlation between the simulated breeding values (TBVs) and estimated breeding values (EBVs).

Accuracy is a function of the heritability of the trait (Canela-Xandri et al., 2015), and the genetic architecture of the trait (Daetwyler et al., 2010; Hayes et al., 2010). For the two models, prediction accuracy, as expected, increased with increasing heritability. Theoretically, the maximum value attainable for prediction accuracy is the square root of the heritability (Canela-Xandri et al., 2015). For the GBLUP model, the accuracies were between 31% and 93% of the theoretical maximum for the traits. The BayesR model produced accuracies between 41% and 46% of the theoretical maximum attainable for the traits.

Bias is the other parameter of interest in prediction analysis, results given in Table 3.2. The two models produced slightly biased EBVs for the low heritability trait. The EBVs obtained for the moderate and high heritability traits using both models are unbiased, with slopes of about 1.

Table 3.2. The prediction analysis of the BayesR model and GBLUP for the 10 replicates of the three heritability classes. The columns are the simulated heritability class, model name, the average accuracy, the average bias. The standard errors of estimates in parenthesis.

| | GBLUP | | BayesR | |
|---|---|---|---|---|
| **Traits** | **Accuracy** | **Bias** | **Accuracy** | **Bias** |
| **h10** | 0.101 (0.015) | 5.452 (1.192) | 0.136 (0.018) | 4.167 (0.747) |
| **h50** | 0.658 (0.004) | 0.965 (0.035) | 0.321 (0.034) | 0.969 (0.066) |
| **h80** | 0.823 (0.002) | 1.025 (0.011) | 0.365 (0.013) | 0.992 (0.027) |

## 3.4 Discussion

This chapter aimed to evaluate the potential of BayesR as a model to adequately provide estimates of complex traits parameters such as the genetic and residual variances and also truly capture the underlying genetic architecture of these complex traits. I conducted a simulation study using three different heritability

classes. I then proceeded to provide a comprehensive analysis and comparison of BayesR and GBLUP to analyse GWA-type datasets of complex traits.

In Bayesian inference, unknown parameters are treated as random variables coming from a specified prior distribution. The posterior distribution of a parameter sums up the information about the parameter conditional on the observation of the data (Held and Bové, 2014). BayesR assumes the prior distribution of marker effects in a GWA analysis is a mixture of four normal distributions. The justification for this prior distribution is that the majority of the SNPs will not be in LD with a QTL and thus will have no effect, while a handful of SNPs will be in LD with a QTL and have an effect. Consequently, the first component of the mixture distribution contains SNPs with zero effects. The remaining three distributions sample SNPs effects that contribute to increasing proportions of the additive genetic variance in traits.

The BayesR model performed well in estimating the heritability of traits spanning three different genetic architectures (Figure 3.5). The three different heritability traits simulated were low, moderate and high heritability traits with heritability of 0.1, 0.5 and 0.8 respectively. The moderate and high heritability traits were simulated such that the distributions had decreasing proportions of the additive genetic variance moving from the second component $k2$ to the fourth component $k4$ of the mixture distributions. The low heritability trait was on the other hand, simulated to have increasing proportions of the additive genetic variance moving from the second to the last distribution. The BayesR model was able to dissect the genetic architecture of these traits (Figure 3.7). On the whole, the model correctly shows traits that have a polygenic architecture, perhaps because the model primarily

assumes polygenicity in the way it defines the second component $k2$ of the mixture distributions to have a large number of SNPs individually contributing a small effect. GBLUP does not offer such partitioning of the genetic variance to indicate how SNPs with effect contribute to the traits.

BayesR detects almost all the SNPs simulated to have the largest effect in the moderate and high heritability traits (Figure 3.6). This is almost twice as good as GBLUP that could only manage to detect half of those on average. The power to detect a SNP as significant in a GWA analysis depends on the variance explained by the SNP. In GBLUP this consequently depends on the LD between the SNP and the causal SNP, the effect of the causal SNP and its frequency (Yang et al., 2010). Depending on the MAF and the effect size, causal variants may explain little variance and thus may not be picked up by GWAS models even though they may be in high LD with genotyped SNPs, however the variance explained by these SNPs aggregate to form part of the total variance explained and thus the GBLUP model will provide good estimates of the additive genetic variance (Figure 3.5). With that being said, LD also presents another problem to GBLUP models such that granted a causal SNP has a large MAF and a big effect and thus explains a large enough variance to be picked by the model and this SNP is in strong LD with multiple genotyped SNPs then its effect may be replicated (Speed et al., 2012). Speed et al. (2012) explore the possibility of this happening in detail and provide an analytical approach to deal with replicated effects of causal SNPs due to LD. The overall additive genetic variance estimate may not be affected by this because the overestimation of the effect in some part of the

genome due to high LD will be compensated for by the underestimation of the effect in other parts of the genome due to low LD.

There was no apparent relationship between MAF and estimated SNP with effects for BayesR (Figure 3.8). Similarly, this was observed for all genome-wide SNPs (Figure 3.9 and Figure 3.10). BayesR provides an estimate of the total number of SNPs that affects a trait. These estimates for the number of SNPs shown in Table 3.1 and the estimates of the additive genetic variance are not very far off the actual numbers simulated for all the traits. But looking at the results shown in Figure 3.6, the SNPs the model estimated to have an effect were only a small minority of the SNPs simulated to have an effect. This means although the estimates are correct, the model picks the wrong SNPs. The question then is, do the results that we get out of the models reflect what was simulated other than the sense that the models pick up something close to what is simulated? The SNPs estimated to have an effect could more or less explain the variance I simulated and I, therefore, decided to investigate if they lie within regions of simulated SNPs (Figure 3.11). This was done by aggregating the SNP effects across the genome in 1000 SNP bins. The plot for the simulated SNP effects appeared in spikes across the genome showing regional clustering of SNPs selected to have an effect with some regions having no effect (Figure 3.11). The plot for the estimated effect, on the other hand, was smoothed across the whole genome with one or two spikes in some regions. This observation may be down to three reasons. The first is that SNPs may be spending more time in the zero-effect distribution (based on the PIPs) than in the distributions with effect over all realisations. The Gibbs sampler at each sampling round estimates the effect

of a SNP and then classifies the SNP based on the effect it has estimated and place the SNP in the appropriate distribution. At the end of the chain, the mean effect across all realisations is calculated for each SNP and if SNPs are sampled most of the time to have no effect, then this will downwardly impact the effect estimate in the end. The second reason is that the washout of the estimated effect observed in Figure 3.11 might be due to the effect of LD: i.e. the split of the effect between all SNPs in LD with the effect SNP. This is because if two SNPs are perfect proxies for each other (i.e. LD of 1) and one is causal, the model may not be able to distinguish between them and at each one sampling stage, the model may give one SNP a posterior inclusion probability of 1 for having an effect and the other SNP a PIP of zero and at the end of the MCMC chain, the effect may be split over both. With that being said, the BayesR model works such that if the SNP effect is not well estimated consistently over realisations, then the model has no way of knowing the correct class for a SNP. The standard error of an estimate gives an indication of how good an estimate is. All SNPs with effect estimates that are less than their standard errors won't be correctly classified by the model and thus may not be picked by the model. The BayesR model's ability to detect effect SNPs is not entirely dependent on LD at least not in the same sense that LD affects GBLUP (as explained above); the BayesR model performs better than GBLUP at correctly picking out SNPs with large effect. The last reason, therefore, for the observations in Figure 3.6 and Figure 3.11 may be that the population size is quite small which will greatly impact the standard errors of the effect estimates. The standard error increases with decreasing sample size. So, although all the SNPs with effect on aggregate explain the total additive genetic

variance, their individual effects may not be picked up in any GWA model, at least not at any genome-wide significance, except for the few large ones. Thus, the effects of background polygenic markers with small effects may be poorly estimated and thus be missed by the models. BayesR still edges GBLUP in its ability to capture the SNPs with large effect. This shows that if there is sufficient information for the BayesR model to act on (for instance large sample size or effect size), it will have sufficient power to correctly pick loci that have an effect on traits. Therefore, increasing the sample size will improve the individual locus effect estimates we get out of this model.

The usefulness of a GWA model can be assessed by its ability to predict unobserved phenotypes of individuals using their genotypes. The accuracy of the prediction reported as the correlation between the simulated breeding values (TBVs) and estimated breeding values (EBVs), is used to assess predictive performance of models. The accuracy of prediction depends on a number of factors, one is SNPs that are in LD with causal loci and also the availability of SNPs that adequately capture the relationship structure between the training dataset and the test dataset (Habier and Fernando, 2013; Habier et al., 2008). Thus, having a lot of related individuals in your data improves accuracy. Nearly a third of the individuals (797) used in the study were close relatives with relationship greater than 3$^{rd}$ degree cousins. Out of these individuals, about 658 had relationships at the level of grandparent – grandchild and above. There was, therefore, sufficient relationship amongst study individuals to provide good prediction accuracies. And this was observed for GBLUP which gave very good accuracies for the moderate and high heritability traits. These accuracies

were better than the ones obtained with BayesR (which do not take advantage of familial relationships in the data) for the same traits. BayesR, however, provides slightly better prediction accuracies than GBLUP for the low heritability traits (Table 3.2).

Besides, family relationships in the data, the size of the training set used in the analysis and the marker density also impacts prediction accuracy (Makowsky et al., 2011). With a training set size of 470,000 individuals, which is the largest human training set ever used, Canela-Xandri et al. (2015) obtained accuracies that were between 68% and 86% of the theoretical maximum. By the deterministic formula for computing the prediction accuracy of additive genomic values (Daetwyler et al., 2008, 2010), my training set size of 2312 individuals and 5520 independent effect loci would generate prediction accuracies of about 0.2005, 0.4161 and 0.5009 for the 0.1, 0.5 and 0.8 heritabilities respectively, not taking family relationships into account. These accuracies will be between 56% and 63% of the theoretical maximum. BayesR produced accuracies that were between 41% and 46% of the theoretical maximum attainable for the traits, which goes to show that BayesR performs very well.

Both models produce unbiased estimates of breeding values for the moderate and high heritability traits, but slightly biased EBVs for the low heritability traits (h10). Thus, for the h10 traits estimated breeding values are lower than the true breeding values. Bias is the slope of the regression of the TBVs on the EBVs. A predictor is unbiased if it has a slope of 1. When a predictor is biased, then the breeding values it estimates are expected to change when more information is accumulated. Therefore, the scale of the EBVs will be stretched towards the TBVs as

more information (more sample) is accumulated in the prediction analysis for the h10 traits.

I will comment on the general applicability of BayesR to real phenotypes by stating that the default implementation that makes use of arbitrary values of $0, 10^{-4}, 10^{-3}, 10^{-2}$ as values for scaling the additive genetic variances for markers assigned to the mixture components works very well and in my case, there was no real advantage in knowing and using the true SNP variances in the model. Also, I further tested the robustness of the BayesR model by changing the starting values for the Dirichlet priors from $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3, \alpha_4) = 1$ to $\boldsymbol{\alpha} = (1,1,1,10)$. This again showed BayesR gives good estimates of the parameters and is not strongly influenced by the starting values of the priors used.

Another point I will make here is that, the phenotypes were simulated based on the assumption of BayesR: the effect of genetic markers are sampled from four normal distributions. This simulation assumption is quite simplistic and does not cover a wide enough range of complex trait architectures as may occur in real phenotypes. And also, the simulation may seem to give BayesR an unfair advantage over GBLUP when comparisons are drawn. An ideal scenario will be to draw additional comparisons between BayesR and GBLUP by simulating phenotypes under the GBLUP assumption: every genetic marker has an equal chance of contributing to the genetic variance. This will essentially mean simulating phenotypes where each genetic marker has an effect on the phenotype. However, the effects of genetic markers for such a phenotype will be completely washed-out across the genome, which will make mapping efforts basically impossible, and also such a phenotype will

not be representative of real-life scenarios. Therefore, although the simulated phenotypes may not be exhaustive enough in terms of the range of phenotype architectures possible and may be biased towards BayesR, it is a good start point. This is because it helps put in context what we should expect to find if indeed real phenotypes are controlled by genetic markers with effects sampled from more than one distribution.

To sum up, I have performed a simulation study that assessed the performance of two GWA models, GBLUP and a Bayesian mixture model. Both models provide good estimates of the additive genetic variation explained by SNPs. The Bayesian model edges GBLUP in capturing SNPs with effect, however, GBLUP provides better prediction accuracies. The Bayesian model presents a unique opportunity to dissect the genetic architecture of traits; that is the number of loci affecting trait variation, their contribution to the additive genetic variance, and distribution of allelic effects. The estimates from the Bayesian model like any other GWA model will improve with the availability of more data. Therefore, although some statistical methods will have better power than others in capturing variants associated with complex traits, increasing the sample size will continue to be the best strategy in the quest to dissect the genetic architecture of complex traits.

# Chapter 4

## 4 Use of a Bayesian mixture model (Bayes R) to investigate the genetic control of urine phenotypes

### 4.1 Introduction

The body's metabolites and proteins are largely excreted through the urine. The urine concentration of electrolytes and proteins, therefore, provides insights into the functioning and metabolic capabilities of various body organs. These electrolytes are used by clinicians in the diagnosis of kidney disease.

Unearthing genetic loci associated with these urine electrolytes will offer insights into the genetics of renal function and into the genetic predisposition of populations to chronic kidney disease.

Bayesian approaches have been presented as better suited to deal with the numerous weaknesses catalogued for conventional GWAS (Meuwissen et al., 2001). In the previous chapter, BayesR was shown in a simulation study to have an edge over GBLUP in mapping loci with large effect leading us to conclude that BayesR is a

better alternative for uncovering genetic markers in traits and for predicting unobserved phenotypes.

In this chapter, I present a genome-wide association study of urine electrolytes using a Bayesian mixture model (Erbe et al., 2012). A total of nine phenotypes comprising of measured urine electrolytes from about 3,000 individuals from the Generation Scotland: Scottish Family Health Study (Smith et al., 2012) were used. This study was designed with the aim of investigating the underlying genetic architecture of these urine phenotypes and to provide estimates of the additive genetic variance explained by the genome-wide SNPs.

## 4.2 Methods

### 4.2.1 Genetic architecture of urine traits

The BayesR model assumes a mixture of four normal distributions with zero means as the conditional priors for the distribution of SNP effects. This assumption is intuitively based on the fact that most of the SNPs will not be in LD with a QTL and thus will have no effect, while a minority of SNPs will be in LD with a QTL and have an effect. As a result, the first component of the mixture distribution contains SNPs with zero effects and the remaining three distributions contain SNPs with non-zero effects. The SNPs in each of the mixture have effects that are drawn from distributions that explain increasing proportions of the additive genetic variance in traits.

The proportion of the additive genetic variance that is explained by each mixture distribution was used to investigate the genetic structure across the urine

phenotypes. Most quantitative traits are assumed to be polygenic such that a large proportion of the total genetic variance is explained by a large number of SNPs each explaining a very small amount of the genetic variance. This polygenic background underlies a few SNPs with moderate to large effects. The BayesR model is designed to be able to capture this structure in traits.

The architecture was explored in two ways. First was the whole-genome architecture which considered how all the genome-wide markers contribute to the total additive genetic variance. This was investigated by considering the proportion of the additive genetic variance explained by all the SNPs that are in each of the mixture distributions.

I also assessed the contribution of SNPs on each autosomal chromosome to the additive genetic variance. This chromosomal contribution to the genetic architecture was determined by calculating the proportion of the additive genetic variance explained by SNPs assigned to a particular mixture distribution on a chromosome. The proportion of genetic variance $\upsilon$ is calculated for chromosome $i$ of a trait as follows,

$$\upsilon_i = \sum_{j=1}^{n} \sum_{k=1}^{4} \rho_k \hat{\sigma}_k^2 / \Gamma, \qquad (4.1)$$

where $n$ is the number of SNPs on the chromosome $i$, $\rho_k$ is the posterior probability of inclusion of a SNP to distribution mixture $k$, $\hat{\sigma}_k^2$ is the additive genetic variance explained by distribution mixture $k$, $\Gamma$ is the sum of the product of the total variance explained by all the SNPs and the posterior probability of inclusion of all SNPs. For $m$ genome-wide SNPs, $\Gamma$ is calculated as follows

$$\Gamma = \sum_{j=1}^{m} \sum_{k=1}^{4} \rho_k \hat{\sigma}_k^2 \quad (4.2)$$

### 4.2.2 Genome-wide evidence for association

The evidence for an association of a SNP with a trait is given by the posterior inclusion probability (PIP) value for that SNP for the four mixture distributions. The PIP for the first mixture distribution provides evidence in support of no association of a SNP with a trait. That is because SNPs sampled for that distribution are assumed to have zero effects. The evidence in support of association, therefore, is one minus the PIP of the first distribution. This value was obtained for all the SNPs and then used to generate association plots.

### 4.2.3 Zoom in around top hit SNPs

I extracted SNPs that flanked top hit SNPs by 500kb on both sides for further analysis to explore the LD structure in those regions, the recombination rates and to discover potential gene candidates within that region.

## 4.3 Results

### 4.3.1 BayesR estimates of parameters

Table 4.1 summarizes the estimates of these parameters for each of the nine urine phenotypes. The heritability values were calculated using estimates of the additive genetic variance and the residual variance. The traits all have low heritability values. The number of SNPs that are sampled into each of the mixture distributions is also reported for each trait. For all the traits, more than 99% of the SNPs were sampled to have no effect on the trait and thus placed in the first mixture distribution.

The number of SNPs sampled to have an effect on the traits varies from just over 2000 for urine creatinine to more than 4,000 for urine potassium.

Table 4.1. The posterior estimates of model parameters by BayesR for the urine traits. The columns show the traits, heritability estimates from GBLUP and BayesR, estimated total number of SNPs having an effect on the trait, estimated number of SNPs for the first, second, third and fourth mixture components.

| Trait | GBLUP | BayesR | | | | |
|---|---|---|---|---|---|---|
| | $h^2$ (%) | $h^2$ (%) | No. SNP | nk1 | nk2 | nk3 | nk4 |
| **Calcium** | 5.76 | 4.16 | 3631 | 551460 | 3286 | 294 | 50 |
| **Chloride** | 1.25 | 5.5 | 2786 | 552305 | 2330 | 405 | 50 |
| **Glucose** | 0 | 4.28 | 2728 | 552362 | 2353 | 309 | 65 |
| **Potassium** | 0 | 4.2 | 4484 | 550607 | 4195 | 246 | 49 |
| **Magnesium** | 10.5 | 4.5 | 2854 | 552237 | 2443 | 354 | 57 |
| **Sodium** | 10.52 | 5.2 | 3579 | 551512 | 3219 | 309 | 51 |
| **Osmolarity** | 4.22 | 5.2 | 2972 | 552119 | 2572 | 346 | 54 |
| **Phosphorus** | 0 | 4.6 | 2788 | 552303 | 2425 | 303 | 60 |
| **Creatinine** | 6.14 | 3.5 | 2283 | 552807 | 1842 | 379 | 62 |

### 4.3.2 The genetic architecture of urine phenotypes

The GBLUP model estimates the total variance explained by all SNPs and provides an estimate of the effect for all SNPs but does not give any indication of how many SNPs truly affect a trait. BayesR, on the other hand, gives estimates closer to the true distribution of SNP effects by specifying a mixture of four normal distributions. The model estimates the variance explained by each component of the mixture distributions in addition to estimating the number of SNPs sampled into each distribution. This is the benefit of the BayesR model – it provides a neat way for studying the underlying genetic architecture of traits.

In this study, the genetic architecture of the traits was investigated first at the whole genome level and subsequently at the chromosome level for the urine traits. The whole genome architecture for a trait was calculated by dividing the additive genetic variance explained by each mixture distribution by the total additive genetic variance.

For estimating the chromosomal architecture, first, the proportion of the additive genetic variance explained by each chromosome was calculated by summing the product of the additive genetic variance contributed by a mixture distribution and the posterior probability of inclusion of SNPs allocated to that mixture distribution per chromosome. This was then divided by the sum of the product of the total variance explained and the posterior probability of inclusion of all SNPs on the chromosome.

The results for the whole genome architecture of traits are shown in Figure 4.1. The results closely mimicked the chromosomal architecture for all the traits; which are urine Calcium (uCa), urine Chloride (uCl), urine Creatinine (uCr), urine Glucose (uG), urine Potassium (uK), urine Magnesium (uMg), urine Sodium (uNa), urine Osmolarity (uO) and urine Phosphorus (uP).

All the traits have a polygenic architecture. This means most of the additive genetic variance is explained by SNPs with small to moderate effects. These SNPs are sampled into the second and third components of the mixture distribution of SNP effects. These two distributions explain a larger proportion of the additive genetic variance of these traits than $k4$. The fourth mixture distribution samples SNPs with

large effects and this distribution explains a lesser proportion of the additive genetic variance in the urine traits.

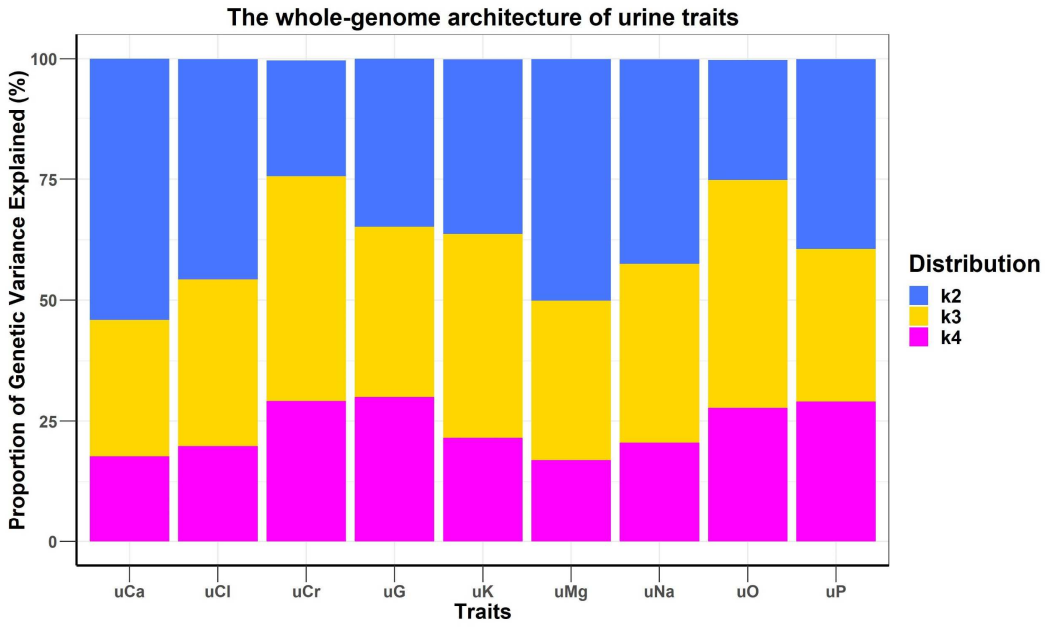



Figure 4.1. The whole-genome architecture of the nine urine traits determined using BayesR. The proportion of additive genetic variance explained by SNPs with effect. The plot bars are colour coded to correspond to the proportions of genetic variance explained by the three mixture distributions that contain the SNPs with effect. The proportions were calculated by dividing the additive genetic variance explained by each mixture distribution by the total additive genetic variance explained by all SNPs. All the traits have an underlying polygenic architecture with mixture distributions that samples small and medium effects SNPs, $k2\ and\ k3$, explaining a larger proportion of the additive genetic variance.

The results for the chromosomal architecture are shown in Figure 4.2 – Figure 4.4. Again, all the traits are polygenic across all the chromosomes. The proportions of variation explained by the chromosomes were mostly driven by the length of the chromosomes (Figure 4.2 – Figure 4.4).

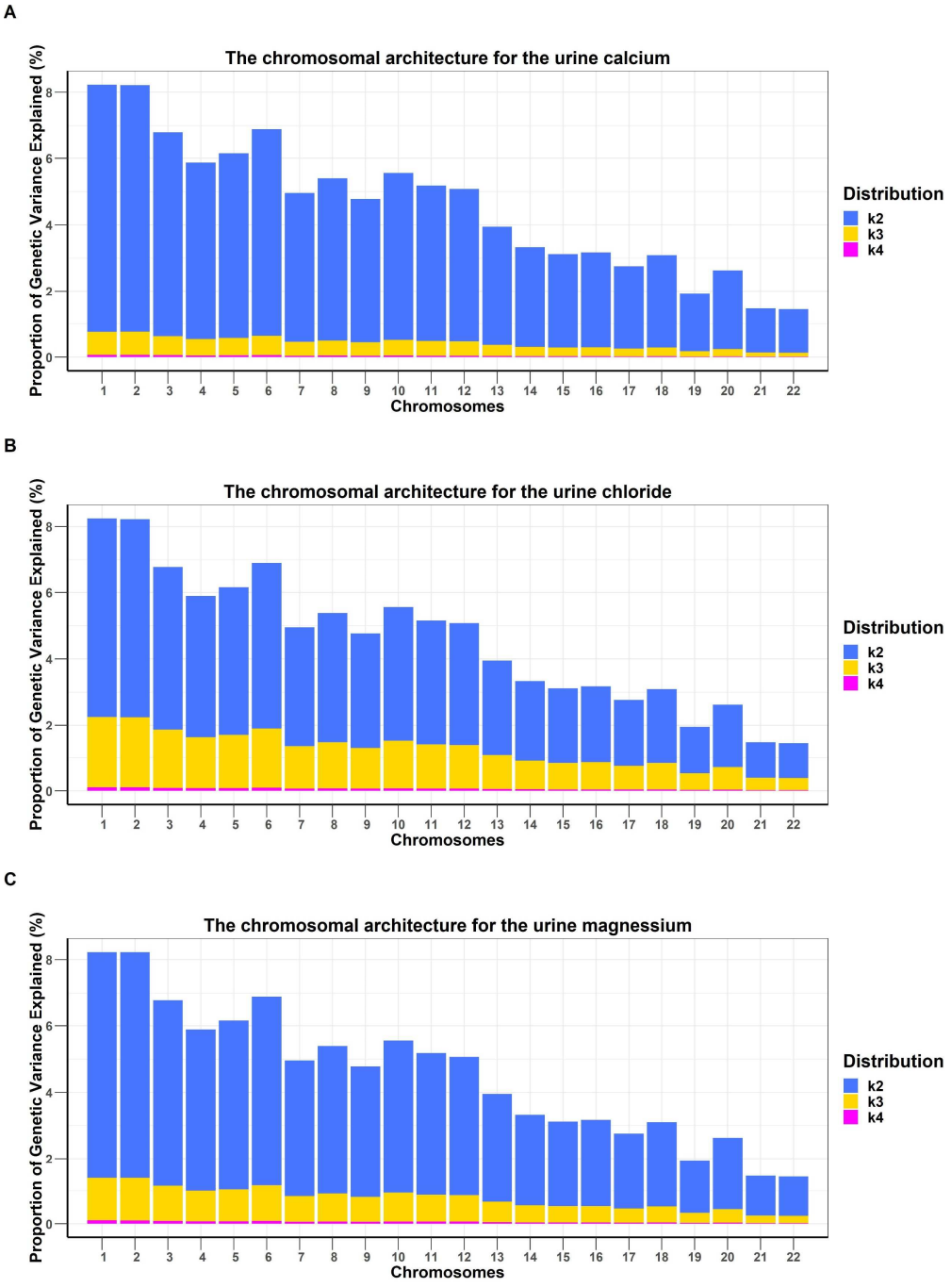



Figure 4.2. The genetic architecture explained by SNPs with effect on each chromosome for urine calcium, chloride and magnesium. The proportion of additive genetic variance explained by each chromosome and the proportion of additive genetic variance explained by SNPs sampled with effect on these chromosomes are plotted. The plot bars are colour coded to correspond to the proportions of genetic variance explained by the three mixture distributions that contains the SNPs with effect.

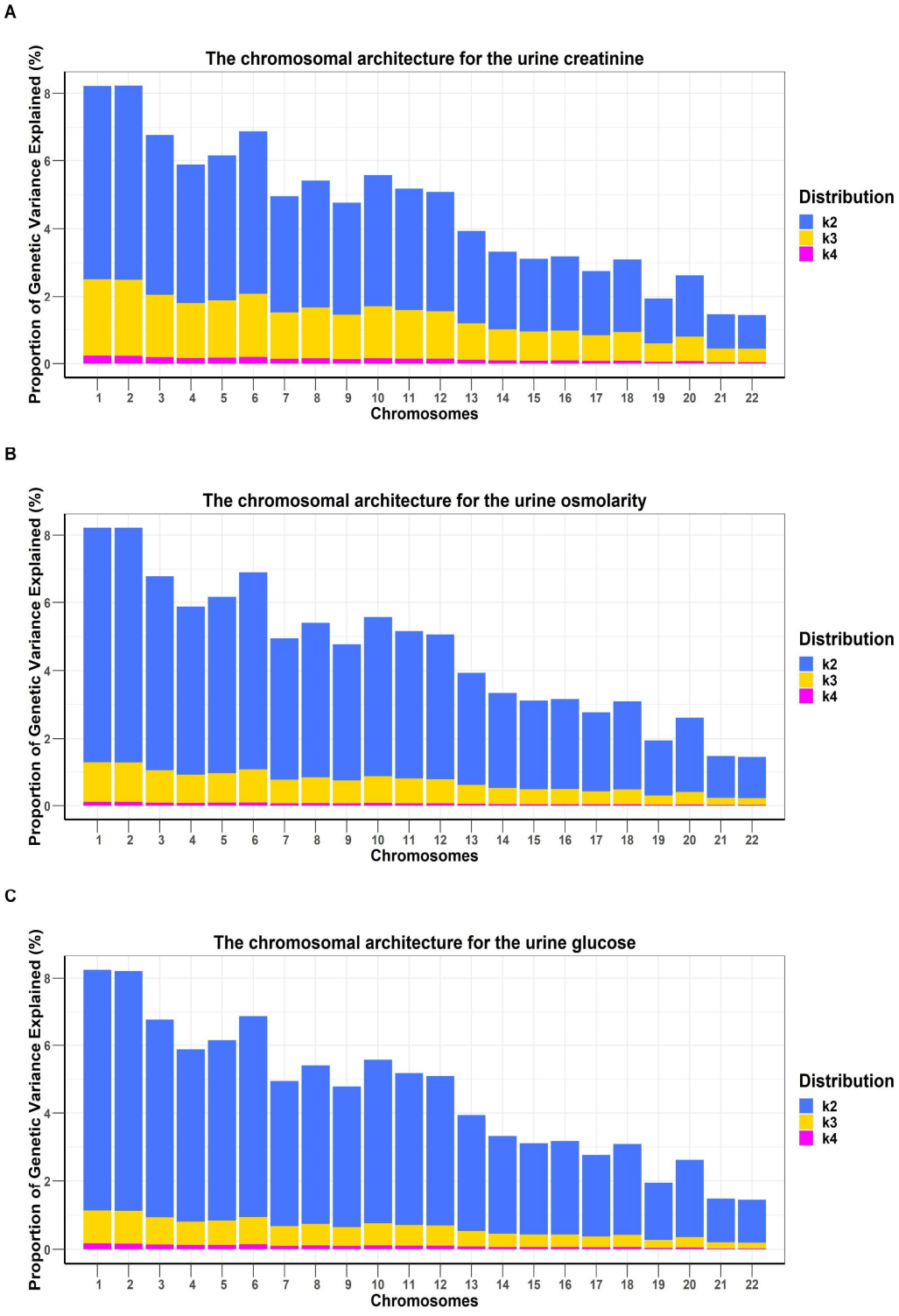



Figure 4.3. The genetic architecture explained by SNPs with effect on each chromosome for urine creatinine, osmolarity and glucose. All the three traits follow a polygenic genetic architecture.

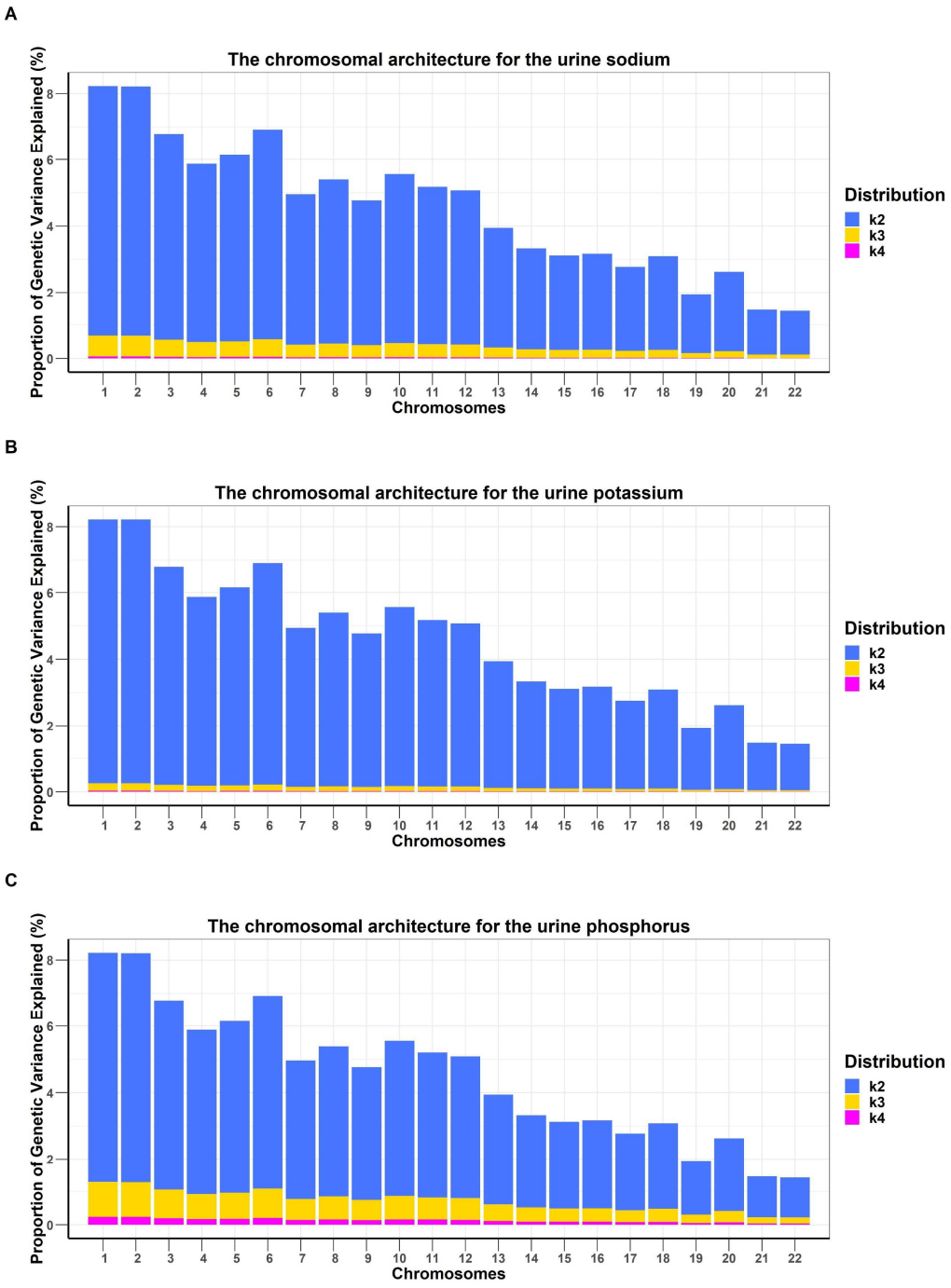



Figure 4.4. The genetic architecture explained by SNPs with effect on each chromosome for urine sodium, potassium and phosphorus. All the three traits follow a polygenic genetic architecture.

### 4.3.3 Genome-wide evidence for association

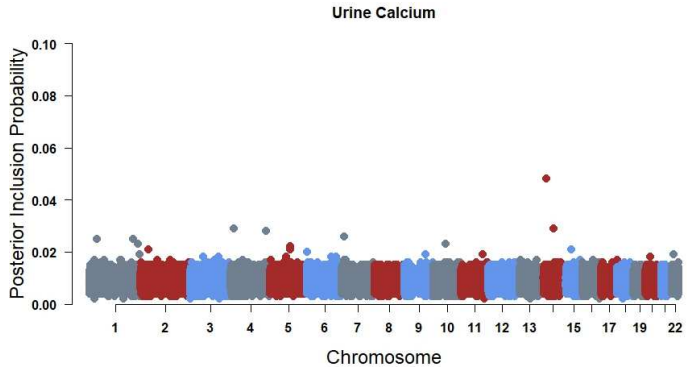
The BayesR model calculates the posterior probability of including (PIP) a SNP in each of the four components of the mixture distribution. The PIPs of the four mixture distributions sum up to unity for each SNP. The first component of the mixture distribution samples SNPs that have no effect on the trait. The PIP of a SNP for this zero-effect distribution is, therefore, evidence of no association of that SNP to the trait. Thus, evidence for association of a SNP to a trait is given by 1 minus the PIP of that SNP for the first distribution. The values obtained from these calculations were used to generate the genome-wide association plots shown in Figure 4.5 – Figure 4.7.

The genome position around the top associated marker for each trait was explored in an analysis to identify genes nearby. Figure 4.8 – Figure 4.11 show the results of this analysis. The genome regions explored are 500Kb upstream and downstream of the top associated SNPs for the urine traits.

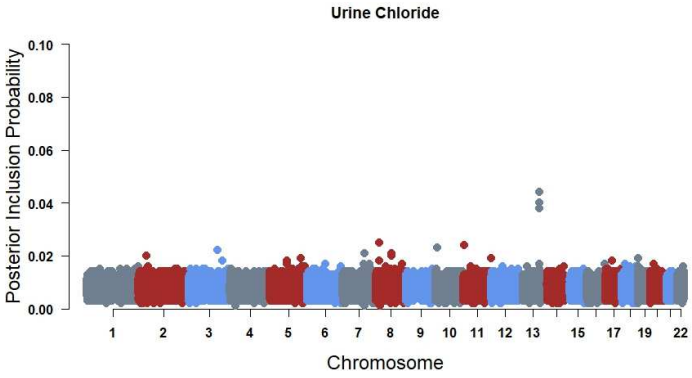
The top SNPs for most traits are located within genomic regions where there are genes close by. Urine creatinine had no gene within the genome region of the highest associated SNP. Table 4.2 summarises the direct gene ontology (GO) class for some of the genes identified to be lying close to the genomic position of the highest associated SNPs for some urine traits. The genes were explored to find possible associations to kidney disease. The highest associated SNP for urine osmolarity, rs795521, was mapped to the *KLF12* gene on chromosome 13 (Figure 4.10). Diseases associated with the *KLF12* gene may include Wegener's Granulomatosis (Wieczorek

et al., 2010), which can cause rapidly progressive glomerulonephritis in the kidney leading to chronic kidney failure.

a.




b.




c.

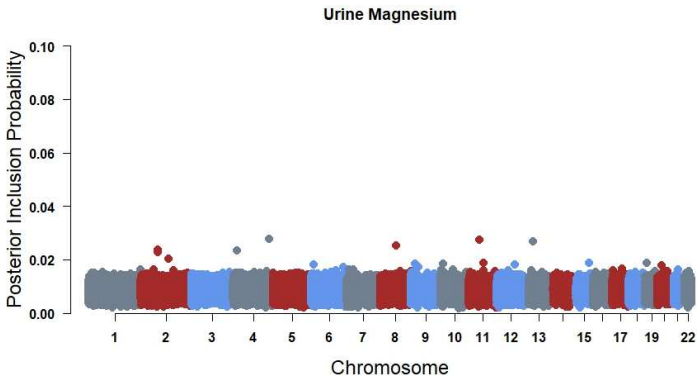
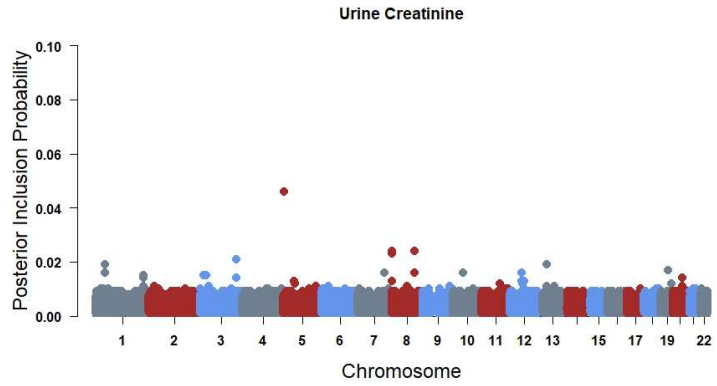
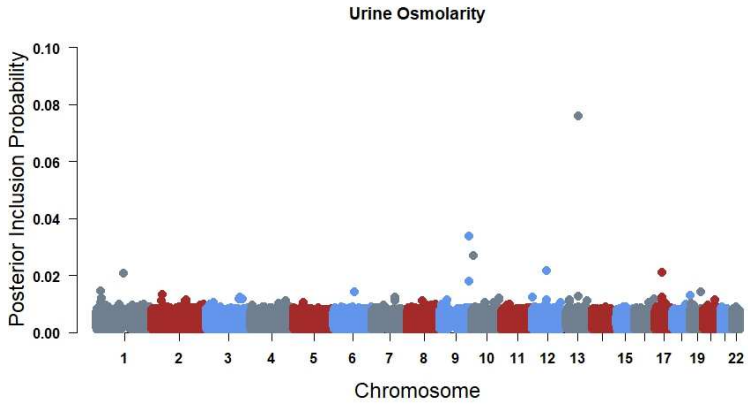



Figure 4.5. The genome-wide evidence for SNP association for urine calcium, chlorides and magnesium. The posterior inclusion probability (PIP) plotted as evidence in support of association for each genome-wide SNP was calculated as one minus the posterior probability of including a SNP in the first component of the mixture distribution. The first component of the mixture distribution is assumed to have zero effect on a trait. The most significant SNP for urine calcium is on chromosome 14, for chloride is on chromosome 13 and for magnesium is on chromosome 4.

a.




b.




c.

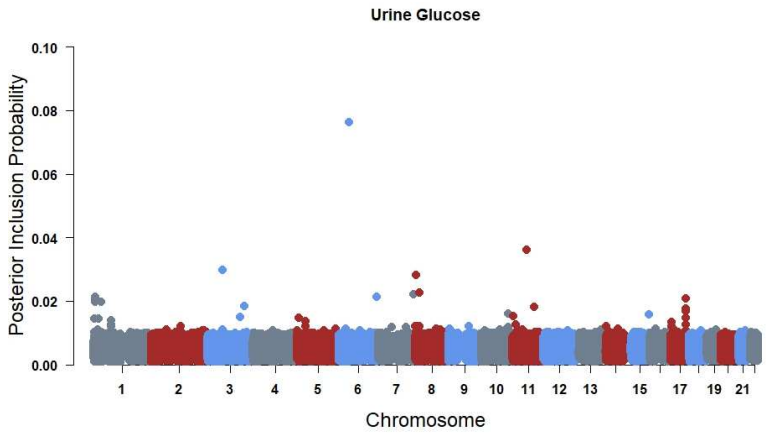
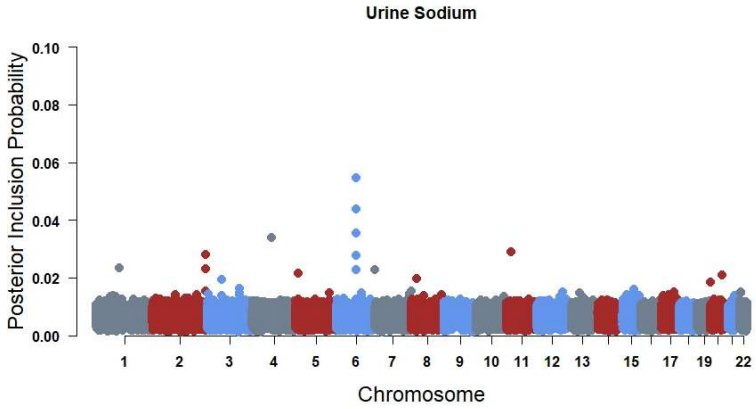
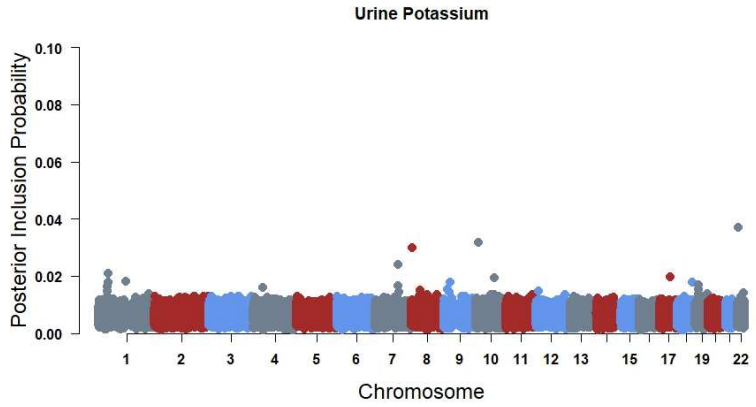



Figure 4.6. The genome-wide evidence for SNP association for urine creatinine, osmolarity and glucose. The most significant SNP(s) for urine creatinine is on chromosome 5, for osmolarity is on chromosome 13 and for glucose is on chromosome 6.

a.




b.




c.

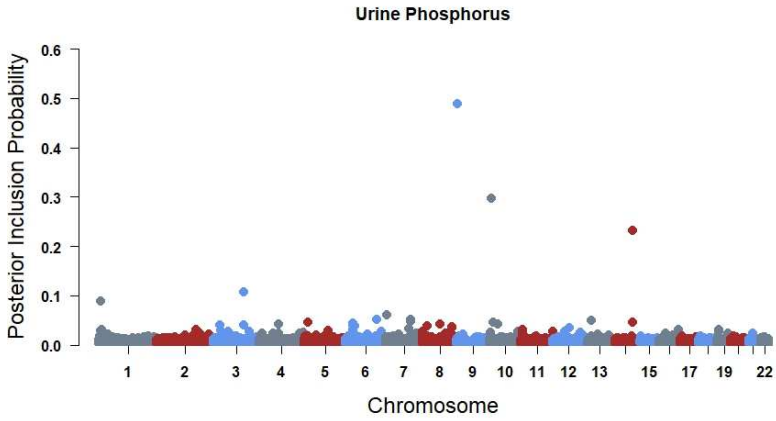
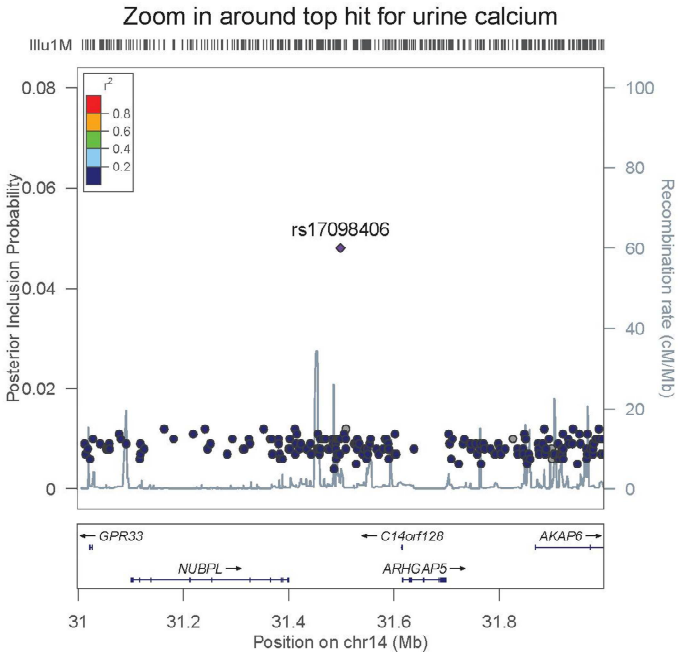



Figure 4.7. The genome-wide evidence for SNP association for urine sodium, potassium and phosphorus. The most significant SNP for urine sodium is on chromosome 6, for potassium is on chromosome 22 and for phosphorus is on chromosomes 9.

a.




b.

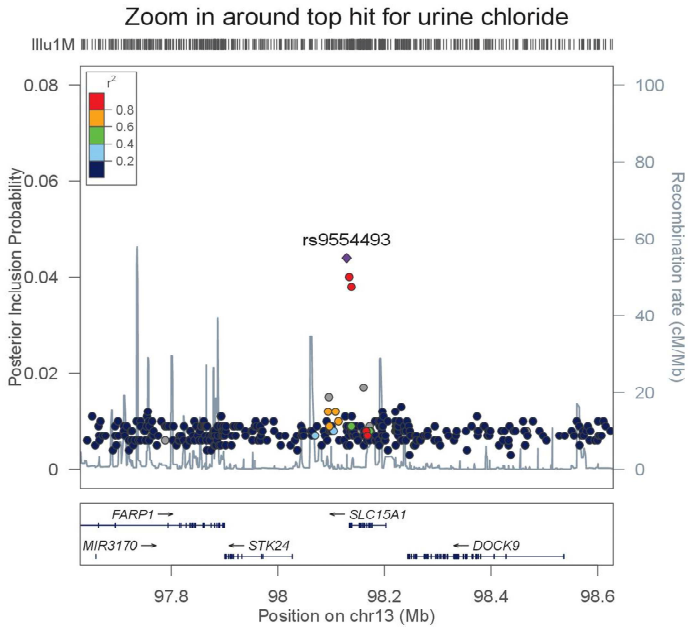
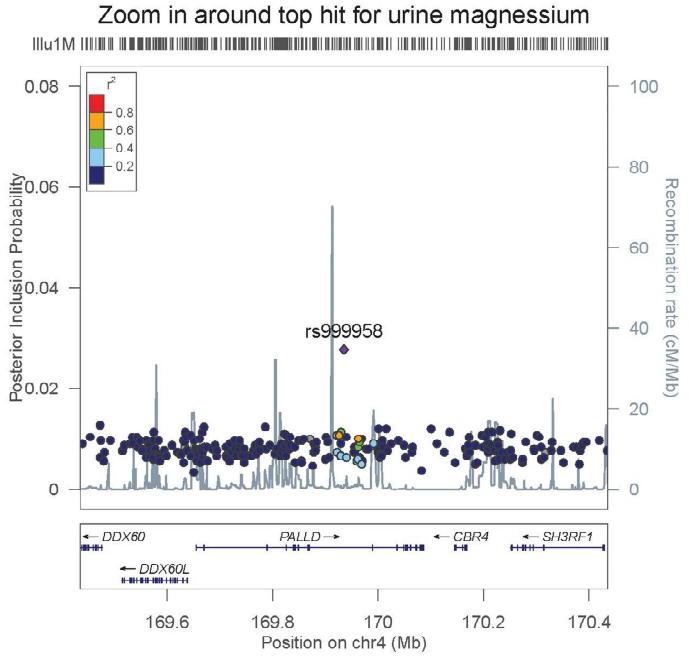



Figure 4.8. Zoom in around 500kb upstream and downstream of the top-hit SNP for urine calcium (a) and chloride (b). The plots are based on the calculation of one minus the posterior inclusion probability of a SNP in the first component of the mixture distribution. The spikes in the plots represents recombination rate within the genomic region. The plot points are SNPs in the region and are colour-coded based on their LD with the top-hit SNP which is shown as a purple diamond. Nearby genes within the genomic region are written at the bottom panels of plots. For urine calcium, the GPR33 and AKAP6 genes have direct GO classes of positive regulation of cytosolic calcium ion concentration and positive regulation of release of sequestered calcium ion into cytosol respectively.

a.




b.

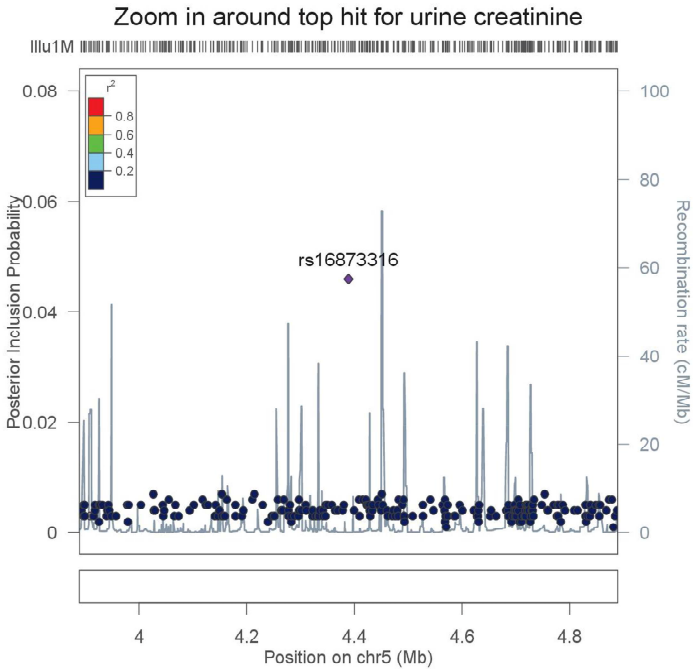


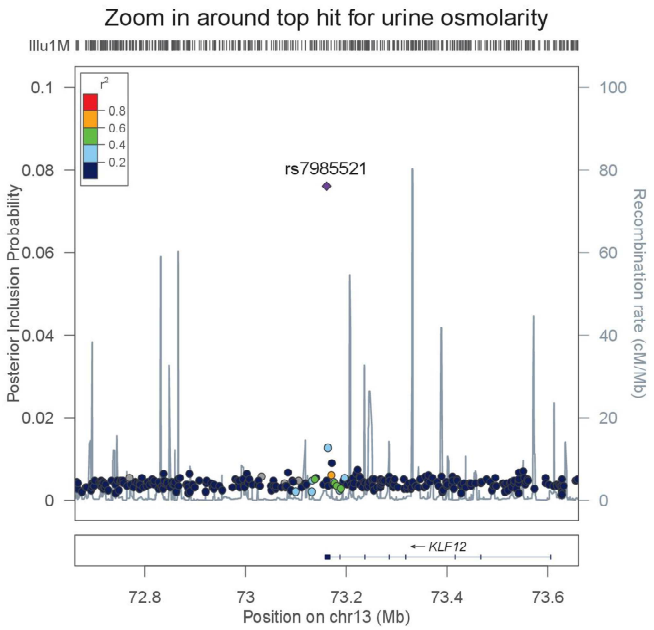

Figure 4.9. Zoom in around 500kb upstream and downstream of the top-hit SNPs for urine magnesium (a) and creatinine (b). The top hit SNP for urine magnesium is in moderate LD with a handful of SNPs within the genomic region and these are bounded by recombination hotspots. The SH3RF1 gene lies downstream of this top SNP and it has a direct gene ontology (GO) class of metal ion binding. The urine creatinine top hit SNP is weakly correlated with other SNPs and there are no nearby genes within the genomic region.

The highest associated SNP for urine sodium, rs7750062, was mapped within 0.32Mb of the *TBX18* gene (figure 4.12). The *TBX18* (T-Box 18) gene is involved in renal development and diseases associated with the gene include congenital anomalies of the kidney and urinary tract 2 and bilateral multicystic dysplastic kidney (Vivante et al., 2015). Congenital anomalies of the kidney and urinary tract are the commonest cause of chronic kidney disease in people within the first three decades of their life (Vivante et al., 2015).

Table 4.2. The Gene Ontology (GO) class for the nearby genes located within genomic regions of top-hit SNPs for urine traits. The columns show the trait name, the gene symbol, the gene name, the direct GO class of gene and the GO reference.

| Urine Trait | Gene | Gene name | GO class (direct) | Reference |
|---|---|---|---|---|
| **Calcium** | GPR33 | Probable G-protein coupled receptor 33 | positive regulation of cytosolic calcium ion concentration | GO_REF:0000033 |
| | AKAP6 | A-kinase anchoring protein 6 | positive regulation of release of sequestered calcium ion into cytosol | GO_REF:0000024 |
| **Potassium** | ZNF280A and B | zinc finger protein 280A and B | metal ion binding | GO_REF:0000037 |
| | GNAZ | Guanine nucleotide-binding protein G(z) subunit alpha | Metal ion binding | GO_REF:0000037 |
| | GUCA2B | Guanylate cyclase activator 2B | Digestion, excretion and body fluid secretion | Reactome: R-HSA-8935690 GO_REF:0000107 |
| | RIMKLA | N-acetylaspartylglutamate synthase A | metal ion binding | GO_REF:0000037 |
| | CLDN19 | Claudin-19 | calcium-independent cell-cell adhesion via plasma membrane cell-adhesion molecules | GO_REF:0000024 |
| **Sodium** | TBX18 | T-box transcription factor TBX18 | renal system development | PMID:26235987 |
| **Osmolarity** | KLF12 | Krueppel-like factor 12 | metal ion binding | GO_REF:0000037 |
| **Magnesium** | SH3RF1 | E3 ubiquitin-protein ligase SH3RF1 | metal ion binding | GO_REF:0000037 |
| **Glucose** | ENPP4 | Ectonucleotide Pyrophosphatase/Phosphodiesterase 4 | bis(5-adenosyl)-triphosphatase activity | PMID:22995898 |

a.




b.

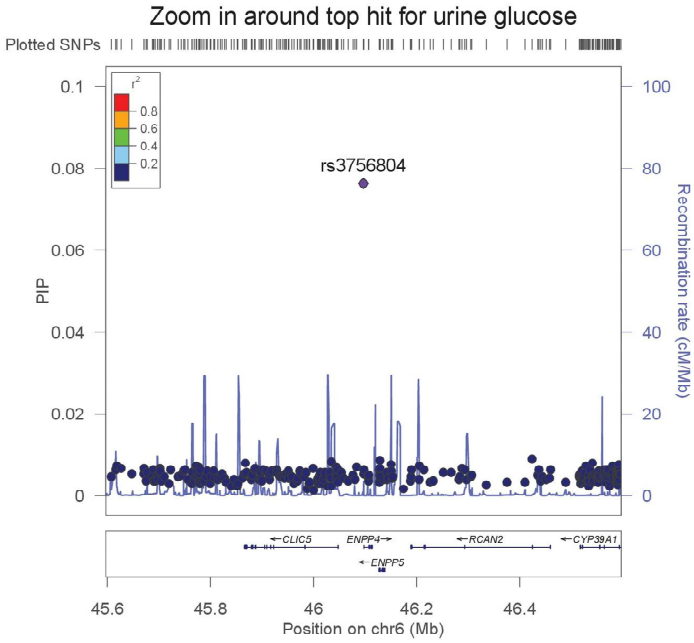
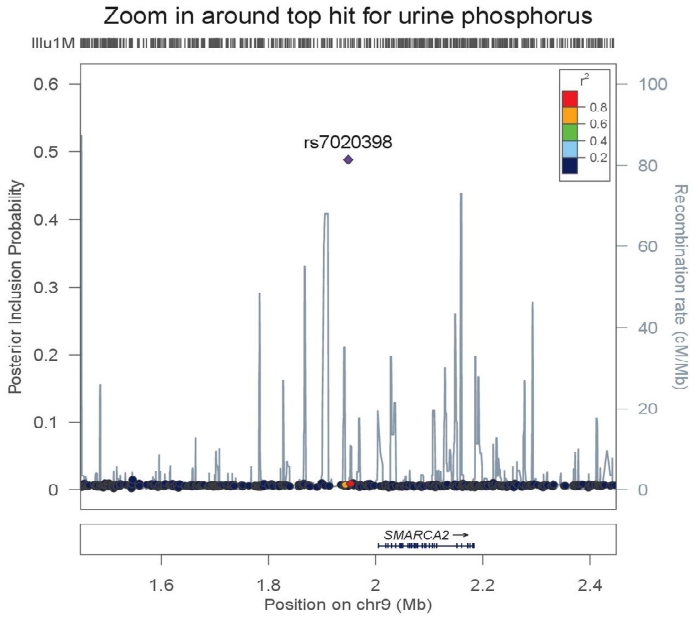



Figure 4.10. Zoom in around 500kb upstream and downstream of the top-hit SNP for urine osmolarity (a) and glucose (b). The top hit SNP for urine osmolarity is moderately correlated with SNPs within two recombination hotspots. The KLF12 gene has a direct GO class of metal ion binding. Diseases associated with KLF12 may include Wegener's Granulomatosis, which can cause rapidly progressive glomerulonephritis in kidney leading to chronic kidney failure. The top SNP for urine glucose is weakly correlated with the other SNPs in the region and that means it is individually driving the genetic association within those regions.

a.




b.

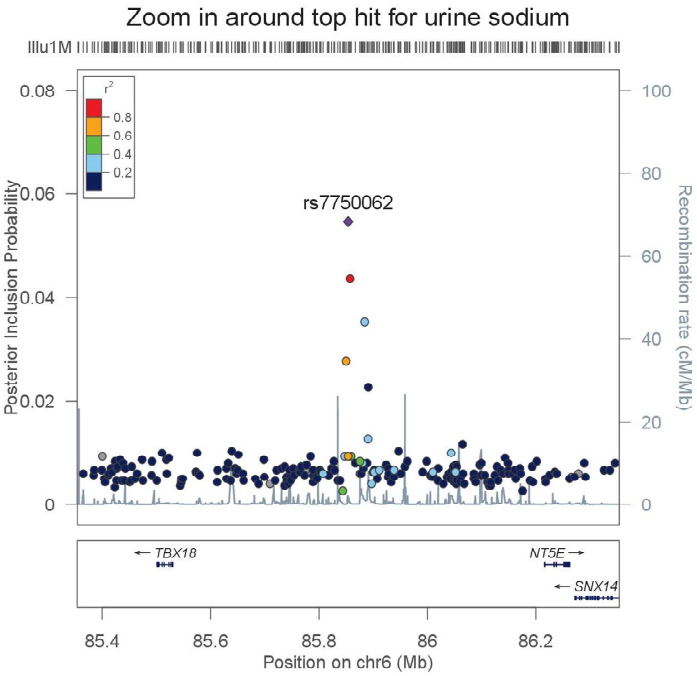



Figure 4.11. Zoom in around 500kb upstream and downstream of the top-hit SNP for urine phosphorus (a) and sodium (b). The top hits of both traits are in high LD with one SNP and is moderately correlated with other SNPs and these are bounded by two recombination hot spots for urine sodium. For urine sodium, the TBX18 gene lies upstream of this top SNP and has a direct GO class of renal system development.

A top hit SNP in urine potassium, rs4660630, was mapped close to the *GUCA2B* (within 0.23Mb) and *CLDN19* (within 0.346Mb) genes on chromosome 1 (Figure 4.12). The *GUCA2B* gene also known as uroguanylin is involved in electrolyte homeostasis (Forte et al., 1996; Kinoshita et al., 1997) and may be implicated in kidney disease (Rahbi et al., 2012). Variations in uroguanylin were associated with urinary volume and sodium and potassium secretion (Guo et al., 2007). The *CLDN19* gene was found to be expressed in renal segments that are mainly involved in paracellular cation transport (Lee et al., 2006) which may suggest their possible involvement in that process. Lee et al., (2006) reported a decreased expression and delocalization of *CLDN19* in polycystic kidneys suggesting a possible link between the gene and kidney disorders.

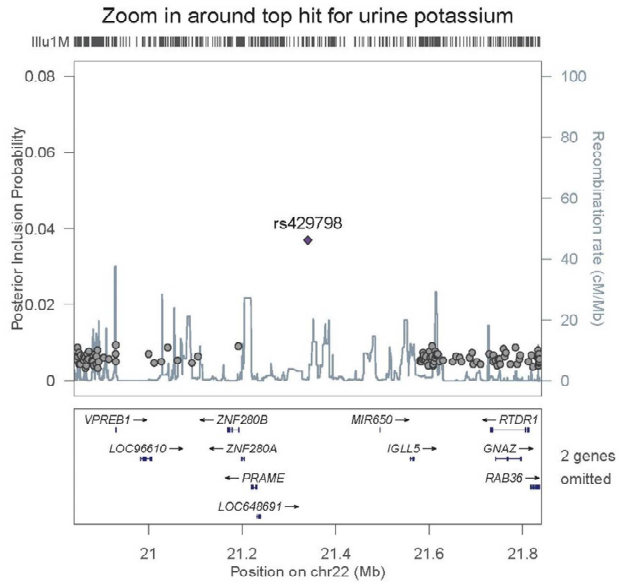
## 4.4 Discussion

I used a Bayesian mixture model (BayesR) to perform a genome-wide association (GWA) analysis on measures of urine electrolytes from about 3,000 individuals from the Generation Scotland: Scottish Family Health Study. The general premise of a GWA analysis is to set measured values of a trait of interest in study individuals as a function of whether an individual has one of three possible genotypes at a genome position. These genotypes are codes of allele counts derived from a genome-wide array of SNPs. The import of a GWA analysis is to test for association between these SNPs and the phenotype of interest.

This can be done in several ways. The most commonly used approach is to fit one SNP at a time in a multiple regression model. The Bayesian mixture model I used

fits all SNPs simultaneously to estimate their effect on the trait. But instead of assuming all the SNPs have an effect on the trait variation and that these effects are

a.




b.

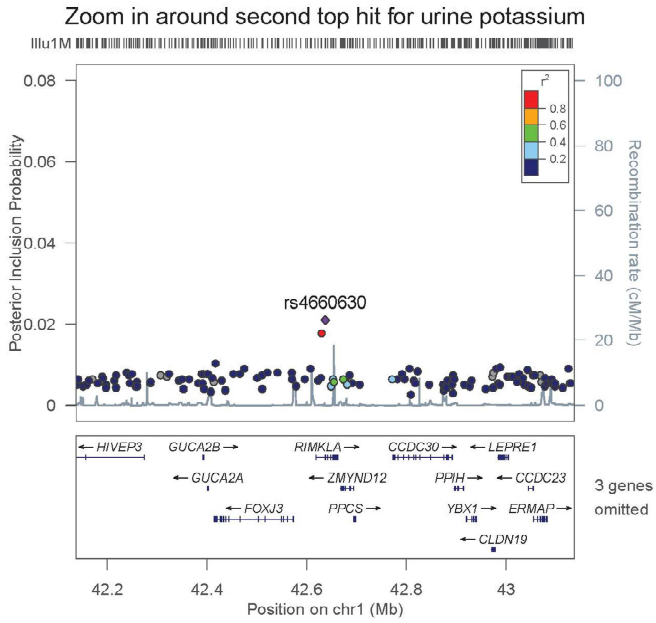



Figure 4.12. Zoom in around 500kb upstream and downstream of the first (a) and second (b) top-hit SNPs for urine potassium. In the lower panel, the top SNP is in high LD with one SNP within two recombination hotspots. There are lots of genes nearby the top hits. The ZNF280A and ZNF280B, GNAZ, and RIMKLA genes have a direct GO class of metal ion binding. The GUCA2B may be implicated in the regulation of salt and water homeostasis in the intestines and kidneys. A decreased expression and delocalization of CLDN19 gene was reported by Lee et al., (2006) in polycystic kidneys suggesting a possible link between the gene and kidney disorders

drawn from the same normal distribution, the model assumes that the majority of the SNPs do not have an effect on the trait and also their effects are drawn from a mixture of four normal distributions with increasing contribution to the total additive genetic variance.

The BayesR model, therefore, provides an estimate of the total number of SNPs having an effect on a trait. It has been established long before GWA analysis that quantitative traits may be controlled by multiple genetic loci and methods that try to map these loci had already been developed (Haseman and Elston, 1972). Subsequently, GWA studies in complex traits, starting with the Wellcome Trust Case Control Consortium (WTCCC) (Burton et al., 2007), confirmed this polygenic architecture for human complex traits and suggested that complex traits may be affected by thousands of genetic loci individually contributing a small effect (Stranger et al., 2011; Visscher et al., 2012, 2017). For the urine traits studied, BayesR estimated the number of effect SNPs to be between 2283 and 4,484 SNPs. This means over 99% of the SNPs tested have no effect on these traits. When the BayesR model was used to analyse binary traits from the WTCCC, over 96% of the SNPs were found to have zero effect on the traits (Moser et al., 2015). These effect SNP numbers obtained from BayesR support the widely held belief that complex traits may be under the control of thousands of genetic loci (Visscher et al., 2017). And the direct implication of these results is that it won't be long for most loci affecting most complex traits to be mapped, since true effect SNPs of most complex traits may be just a few thousand loci. These loci-mapping efforts will be greatly aided by the recent advancement in the field of whole genome sequencing, and improved

analytical methods and also significantly increased study sample size such as the UK Biobank (Sudlow et al., 2015).

The BayesR estimates of heritability obtained for the urine traits were generally low but not zero as were obtained for some of the urine traits using GBLUP. Urine glucose had moderate heritability with much of the additive genetic variance explained by SNPs sampled in the large effect component of the mixture distribution.

All the urine traits have a polygenic architecture (Figure 4.1). This polygenic architecture was also observed in the chromosomes for the traits. The proportions of additive genetic variance explained by the chromosomes were positively correlated with chromosome size in all traits (Figure 4.2 – Figure 4.4).

Figure 4.5 to Figure 4.7 show that the BayesR model can identify SNPs that have effects on traits, but often the evidence of association (1 minus PIP of k1 component of mixture distribution) was low but not zero. The overarching aim of a GWA analysis is to understand the underlying biology of traits, especially diseases, in the hope of gaining a better understanding of disease aetiology which may translate to improved prevention and treatment. To this end, I mapped the highest associated SNPs to genomic regions to identify genes close by and explore possible links between these nearby genes and the urine traits that have been measured in my study population.

All the highest associated SNPs for the urine traits, except urine creatinine, lie in genomic regions which harbour genes (Figure 4.8 to Figure 4.12). One may expect SNPs that are associated with complex traits to lie within genes, but there is also

enough information to suggest that a lot of the variants affecting complex traits reside outside genes in regulatory sequences that control gene expression (Pai et al., 2015). This was the case for most of the highest associated SNPs for the urine traits. The SNPs reside either within genes or in intergenic sequences within the genomic regions. For some of the traits, there were possible functional links between the genes nearby associated SNPs and the urine traits tested. For instance, one associated SNP in urine potassium is situated close to the *CLDN19* gene which has been implicated in kidney disease related phenotypes (Lee et al., 2006). These genes and others identified for the other urine traits could serve as candidate genes in future efforts aimed at finding gene targets for kidney disease therapy.

I conclude by saying that, BayesR is a good analytical tool for studying the genetic architecture of complex traits. The method gives an estimate of the number of loci affecting a trait which potentially becomes useful to efforts in trying to map causal loci of complex traits. The BayesR genome-wide evidence of association given by the posterior inclusion probabilities was mostly low for a lot of the urine traits. This indicates that BayesR like many other analytical methods is susceptible to low power due to the small sample size. The results from BayesR analysis, therefore, can be improved with a larger study sample size.

# Chapter 5

## 5 Regional Heritability analysis of complex traits using haplotype blocks defined by natural recombination boundaries

### 5.1 Introduction

Before the introduction of SNPs in genetic research, the best linear unbiased predictor (BLUP) was the standard method for genetic evaluation, particularly in livestock research (Robinson, 1991). BLUP made use of a relationship matrix that contained identity by descent (IBD) coefficients, calculated using pedigree information of study individuals. The BLUP analytical method was later implemented in a genome-wide analysis setting using an identity by state (IBS) relationship matrix that was calculated using genome-wide SNPs (Meuwissen et al., 2001; VanRaden, 2008). This genomic implementation of BLUP was named GBLUP. Efficient ways for calculating these genomic predictors were proposed by VanRaden (2008), who proposed two genomic relationship matrices (GRMs) in which the covariance of genomic values between two individuals are scaled to be analogous to the pedigree-based relationship matrix used in BLUP.

These GRMs by VanRaden (2008) became widely accepted by genetic researchers and used in a restricted maximum likelihood method known as GREML to calculate the total genetic variance or the heritability (Yang et al., 2010). The GREML method played a pivotal role in uncovering some of the so-called missing heritability of complex traits (Clarke and Cooper, 2010; Maher, 2008; Manolio et al., 2009; Speed et al., 2012; Yang et al., 2011). More than half a decade and several analytical methods later (Chen et al., 2015; Gonzalez-Recio et al., 2015; Hemani et al., 2013; Reynolds and Finkel, 2015; Yang et al., 2015; Zhu et al., 2015; Zuk et al., 2012), the heritability was believed to be hidden (Yang et al., 2015) rather than missing.

The heritability may be missing or hidden, and in both cases, this points to one fact: that current analytical approaches in genetic variance estimation can account for only a fraction of the total contribution of genetic factors to the variation observed in phenotypes. There is, therefore, the need for continual efforts in uncovering the total heritability.

One of the numerous arguments made to account for the missing heritability is that true causal variants of traits may be rare (Pritchard, 2001) and thus may be in incomplete linkage disequilibrium (LD) with genotyped SNPs (Yang et al., 2010). There is, therefore, some benefit to be gained in terms of improving the heritability estimates and uncovering gene variants involved in the control of traits by fitting GWA models that adequately capture rare genetic variants (Cirulli and Goldstein, 2010; Gonzalez-Recio et al., 2015).

I am proposing a genome-wide analytical approach that draws its theoretical basis from a variant of the GREML approach that uses both local and genome-wide relationship matrices to provide regional estimates of the heritability across locally defined regions in the genome (Nagamine et al., 2012). What is unique in this approach is that it utilises a relationship matrix that is based on local haplotype blocks defined by recombination boundaries in the genome.

Compared with SNPs, haplotype analysis has an advantage because haplotypes can be functional units (Vormfelde and Brockmöller, 2007) and thus haplotype analysis can capture the joint effects of closely linked cis-acting causal variants (Balding, 2006). Haplotypes provide a better strategy in capturing true genomic relationship amongst individuals in the presence of rare variants and thus should provide real benefit over SNPs in recovering much of the missing heritability and identifying novel trait-associated variants. This is because although rare variants are not in LD with genotyped variants and thus difficult to capture in GWAS, these rare variants may be in LD with some haplotypes and thus can be captured using haplotype methods.

I hypothesize that this approach will complement already existing GWAS analytical approaches by capturing regions in the genome contributing to the phenotype that existing GWAS methods fail to capture. In this chapter, I report the implementation of this approach on simulated data and explored its performance in detail. Results from the simulation study support my hypothesis and I have confirmed there are real benefits to be gained from this approach by applying it to real data

obtained from about 20,000 individuals from the Generation Scotland: Scottish Family Health Study (GS: SFHS).

## 5.2 Methods

### 5.2.1 SNP-based regional GREML model

The general statistical setting of a regional GREML analysis has been described already in the methods section in chapter 2. This type of SNP-based regional GREML analysis was first reported by Nagamine et al. (2012). The regional GREML analysis approach I employ here differs from the analysis done by Nagamine et al. (2012) in the way the regions are defined. Their analysis defined local regions by breaking the genome into smaller user-defined windows of $n$ SNPs, which overlapped by $x$ SNPs. My model, however, defines local regions naturally based on recombination boundaries in the genome.

The regional GREML model fits two genetic relationship matrices (GRMs): one local GRM for the region and a whole genome GRM for the remaining SNPs in the genome that are outside the region. Both GRMs are kinship matrices calculated as the proportion of the genome-wide autosomal SNPs shared identity by state (IBS) between pairs of individuals. The SNP IBS matrices are calculated as follows, following the second scaling factor proposed by VanRaden (2008)

$$G = \frac{ZZ'}{m} \tag{5.1}$$

where $m$ is the total number of local or genome-wide autosomal SNPs, and $\boldsymbol{Z}$ is a matrix of genotype codes for the sampled individuals which has been centred by loci

means and normalised by the standard deviation of each locus. $\boldsymbol{Z}$ is calculated as follows for individual $i$ at locus $j$

$$Z_{ij} = \frac{(x_{ij} - 2p_j)}{\sqrt{2p_j(1-p_j)}} \tag{5.2}$$

where $x_{ij}$ is the genotype code at locus $j$ for individual $i$ and takes the values 0, 1 and 2 for AA, Aa and aa genotypes respectively, $p_j$ is the frequency of allele 'a' at locus $j$**.** The SNP-based kinship for individuals $i$ and $k$ is therefore calculated as follows

$$G_{ik} = \frac{1}{m} \times \sum_{j=1}^{m} \frac{(x_{ij} - 2p_j)(x_{kj} - 2p_j)}{2p_j(1-p_j)} \tag{5.3}$$

### 5.2.2 Haplotype-based regional GREML model

The haplotype-based regional GREML model follows on theoretically from the SNP-based analysis and utilises haplotypes instead of SNPs as the genetic markers for the local analysis. The analysis fits two GRMs, a haplotype-based regional GRM and a SNP-based genome-wide GRM. The haplotype-based GRM is similar to the SNP-based GRM defined in the previous section. For a locally defined region containing $h$ haplotypes, the haplotype-based kinship for individuals $i$ and $k$ is calculated as follows

$$H_{ik} = \frac{1}{h} \times \sum_{j=1}^{h} \frac{(d_{ij} - 2p_j)(d_{kj} - 2p_j)}{2p_j(1-p_j)} \tag{5.4}$$

where $d_{ij}$ is the diplotype code (coded as the number of copies of haplotype $j$) for individual $i$ and takes the values 0, 1 and 2 for the $h_t h_t$, $h_t h_j$, $h_j h_j$ diplotypes respectively where $t \neq j$, $p_j$ is the haplotype frequency for haplotype $j$.

### 5.2.3 Phenotype simulations

Five phenotypes were simulated using available genotypic information of approximately 20,032 individuals from the Generation Scotland: Scottish Family Health Study (GS: SFHS) (Smith et al., 2012). A total of 593,932 genotyped SNPs were used, and missing genotypes were filled in by imputed data. About 555,091 SNPs remained after a QC that removed SNPs of MAF < 0.01 and SNPs that were out of Hardy-Weinberg equilibrium at p-value < 1e-5.

The five phenotypes were simulated to have a total variance of 1. This is composed of 0.6 environment variance and a genetic variance of 0.4. The genetic variance was partitioned into two components, a polygenic variance of 0.3 and a QTL variance of 0.1. A common polygenic variance was simulated for all five phenotypes from 20,000 markers randomly selected across the genome. Half of the 20,000 markers were randomly assigned negative effects and the other half were randomly assigned positive effects. The polygenic SNP effects were assumed to be normally distributed with zero mean and variance equal to the polygenic variance 0.3.

For each phenotype, 20 regions or haplotype blocks were randomly selected, one on each chromosome (except chromosomes 6 and 8 because of the unusually high LD in the MHC regions on chromosome 6 and inversion duplication regions on chromosome 8), to simulate quantitative trait loci (QTL). This gave a total of 20 QTLs for each phenotype. The regions were delimited by natural boundaries which are recombination hotspots where the estimated recombination frequency do not exceed 10 centiMorgans per Megabase (10cM/Mb) based on the Genome Reference Consortium Human Build 37 (International Human Genome Sequencing Consortium,

2004). The number and type of marker used to simulate the QTL are what defined the five phenotypes. The five phenotypes are, a 1-SNP QTL within the haplotype block, a multiple-SNP (5 SNPs) QTL within the haplotype block, two types of 1-haplotype QTL within the haplotype block and multiple (5) haplotype QTL within the haplotype block. These are described below.

For the haplotype QTL phenotypes, a haplotype block is treated as a single genetic locus having multiple alleles. Each haplotype within a block is considered as an allele of that locus. Each study individual will carry two alleles or diplotypes for each locus or haplotype block. The genotype data used to simulate the phenotypes was phased using SHAPEIT (Delaneau et al., 2012) and haplotypes for study individuals were extracted. The multiple haplotype QTL phenotypes were simulated by randomly sampling two rare haplotypes and three common haplotypes within each haplotype block to give a total of five haplotypes per block. The two types of 1-haplotype QTL phenotypes were simulated by randomly sampling a rare haplotype per haplotype block for one type and for the other type a common haplotype was randomly sampled within each haplotype block.

The polygenic effect and the QTL effects were calculated as follows

$$\sigma_j^2 = 2\sum_{j=1}^{n} p_j g_j^2,$$

$$g_j = \sqrt{\frac{\sigma_j^2}{2\sum_{j=1}^{n} p_j}}, \qquad (5.5)$$

where $\sigma_j^2$ is the contribution of a QTL to the polygenic variance, $g_j$ is the effect of a SNP $j$ or haplotype $j$ randomly sampled to have polygenic or QTL effect, $p_j$ is the frequency of haplotype $j$ or the effect allele of the SNP $j$, $n$ is the number of alleles at a genetic locus. For a diallelic locus such as those for genotyped SNPs, $n = 2$. Each QTL explained a variance of 0.005. For the single QTL phenotypes, each QTL marker had a variance of 0.005. The multiple QTL phenotypes also had a variance of 0.005 at each locus, and each QTL marker explained a variance of 0.001.

Common environmental effects were randomly sampled for the five phenotypes from a normal distribution $N(0, \sigma_e^2)$ where $\sigma_e^2$ is 0.6. This together with a genetic variance of 0.4 gave a total variance of 1 for each phenotype. The final simulated phenotype for an individual $i$ was then calculated as follows

$$y(single\ QTL)_i = \sum_{j=1}^{20000} x_{ij} g_j + \sum_{j=1}^{20} x_{ij} g_j + e_i, \quad (5.6)$$

$$y(multiple\ QTL)_i = \sum_{j=1}^{20000} x_{ij} g_j + \sum_{l=1}^{20} \sum_{j=1}^{5} x_{ij} g_j + e_i, \quad (5.7)$$

where $x_{ij}$ is the number of copies of haplotype $j$ or the effect allele of SNP $j$ and $g_j$ is the effect of haplotype $j$ or SNP $j$. Twenty replicates were analysed for each of the five phenotypes with a different set of QTL markers sampled for each replicate.

### 5.2.4 Model implementation

The regional GREML model is a mixed effects model that fits both fixed and random effect terms. In this simulation study, the mean is fitted as a fixed effect and

the polygenic, QTL and residual terms are fit as random effects. The model fits two GRMs to account for the local QTL variance and whole genome polygenic variance.

The five phenotypes were analysed using two models, a SNP-based regional GREML model (Sbm) (for the SNP QTL phenotypes) and haplotype-based regional GREML model (Hbm) (for the haplotype QTL phenotypes). To test for the specificity of the analytical models, I applied the haplotype-based regional GREML model to SNP QTL phenotypes and the SNP-based regional GREML model to the haplotype QTL phenotypes. I also performed an Hbm analysis in which the natural haplotype block sizes were restricted to 20 or fewer SNPs per haplotype block. This was to investigate whether the regional effect will be well captured by the haplotype-based model when shorter haplotypes are used.

The SNP-based model GRMs were calculated using the REACTA software (Cebamanos et al., 2014). The haplotype-based model GRMs were calculated using a locally written Fortran programme. The GRMs were then utilised in REACTA to estimate the regional genetic variance and polygenic variance using a restricted maximum likelihood. For each phenotype, I analysed 220 regions in total to map the 20 simulated QTLs. This involved analysing the region containing the QTL and 10 adjacent regions (five in either direction).

## 5.3 Results

I performed a regional GREML analysis that fits two GRMs (one for the region and one for the rest of the genome) per region across multiple genomic regions defined by recombination hotspots. I tested two types of regional GREML models on

20 replicates of five simulated phenotypes. One model fitted a regional SNP GRM (SNP-based model) and the other fitted a regional haplotype GRM (Haplotype-based model), each together with a genome-wide SNP GRM. The phenotypes were simulated to have 20 regional QTL effects and genome-wide polygenic effects. The regional QTL effects of the five phenotypes were simulated using SNPs as causal variants for two of them and haplotypes for the remaining three.

The likelihood ratio test (LRT) was used to test the hypothesis, $H_0$: that the genetic variance explained by the region is not significant, against the alternative, $H_1$: that the region accounts for a significant proportion of the genetic variance. A large LRT statistic is an evidence against the null hypothesis, and therefore means the region explains a significant proportion of the genetic variance.

The LRT averaged over the 20 replicates of the five phenotypes are shown in Figure 5.1 and Figure 5.2. Figure 5.1 shows the average LRT for the QTL regions and 10 adjacent regions. The results show that both models could estimate the simulated regional effects with statistical significance and therefore can capture true causal loci in traits. The LRTs were higher on average for the SNP-based model (Sbm) compared to the haplotype-based model (Hbm).

The models, however, fail to capture the simulated regional effects when the simulated phenotype does not match the analysis model (Figure 5.3a and 5.3b). These figures show the results for the situation where the SNP QTL phenotypes were analysed with the haplotype-based model and the haplotype QTL phenotypes were analysed with the SNP-based model. Both models fail in such situations. Figure 5.3a

and 5.3b show that the models are complementary to each other since they capture effects due to different types of genetic variants (i.e. tagged by SNPs or haplotypes).

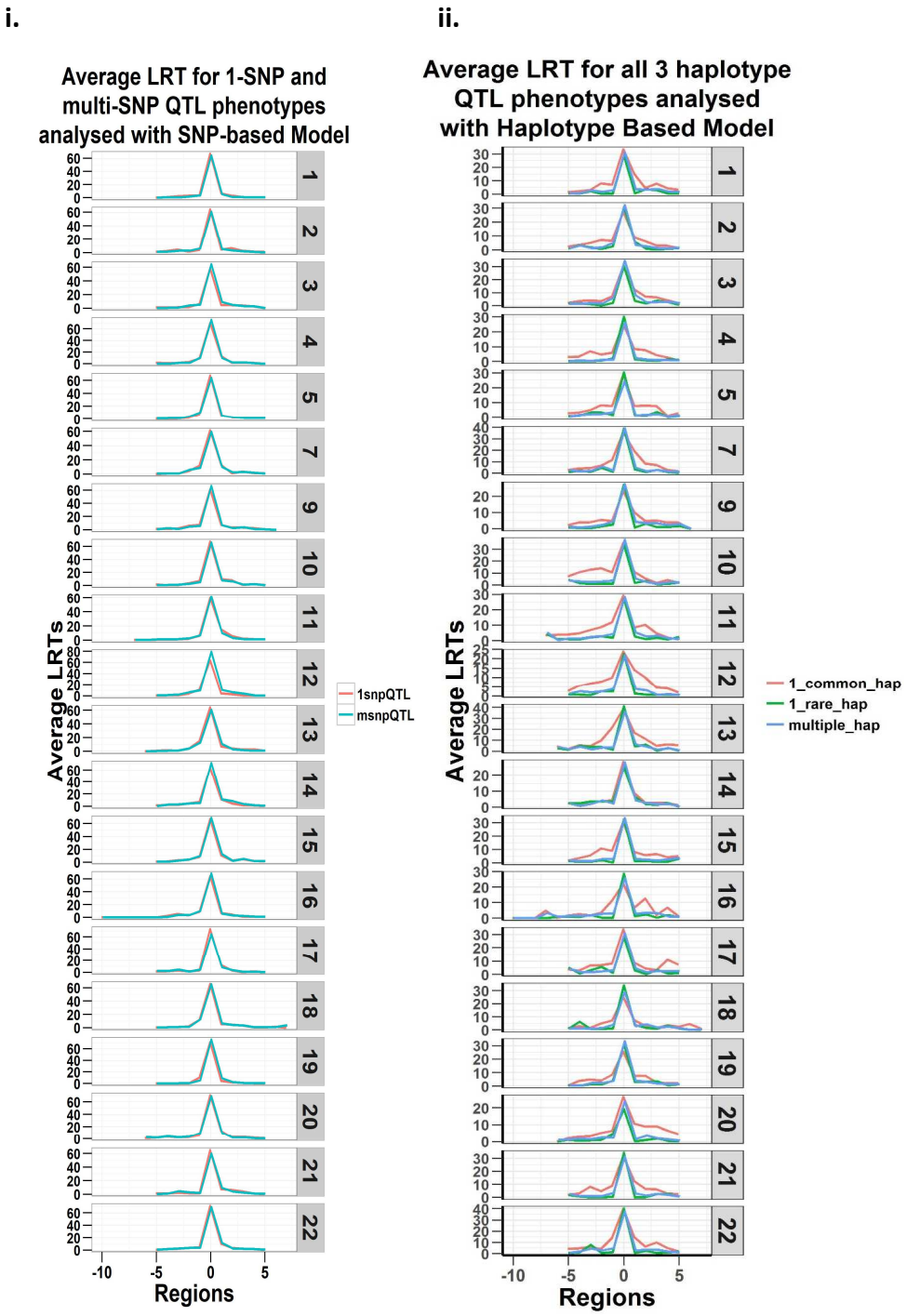
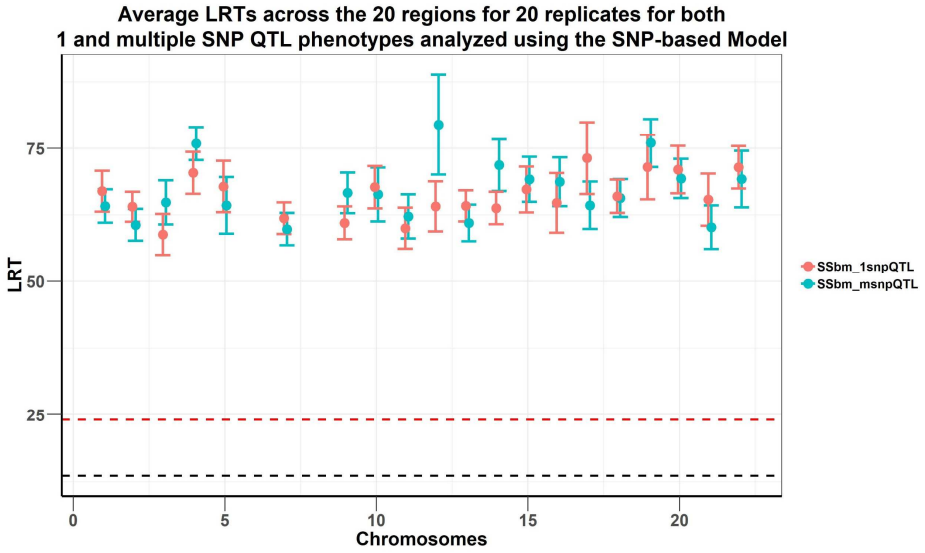



Figure 5.1. Plots of Likelihood ratio test (LRT) statistics at each QTL loci and 10 adjacent regions averaged for the 20 simulations of each of the five QTL phenotypes. Plot (i) is SNP QTL phenotypes analysed using the SNP-based model and plot (ii) is the haplotype QTL phenotypes analysed using the Haplotype-based model. Both models can capture the simulated QTL effects for their respective phenotypes.

The LRT statistics decayed with increasing region size (the number of markers in the region) for the two models even when the models matched the simulated phenotypes (Figure 5.4a). The rate of decay was more pronounced in the haplotype-based model (Figure 5.4b).

i.




ii.

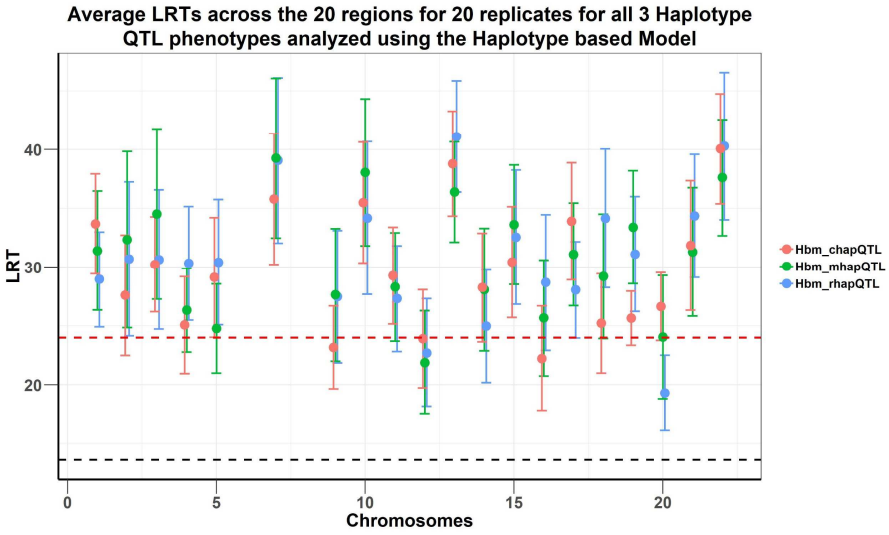
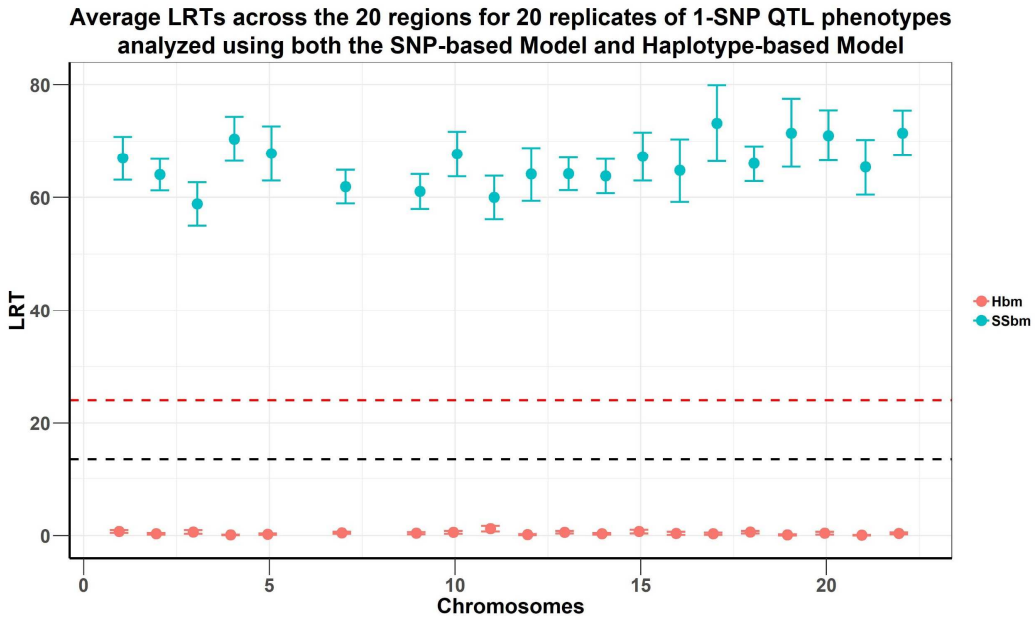



Figure 5.2. Plots of average LRT statistics over replicates of QTL loci across the chromosomes for the 20 simulations of each of the five QTL phenotypes. The red dashed lines are genome-wide significance at alpha of 0.05 and the black dashed lines are Bonferroni significance threshold (for 220 regions). The upper panel (i) is plot of SNP QTL phenotypes analysed using the SNP-based model and the lower plot (ii) is a plot of haplotype QTL phenotypes analysed using the Haplotype based model. Both models, on average, captures the effects SNP and haplotype QTLs at loci very well.

i.




ii.

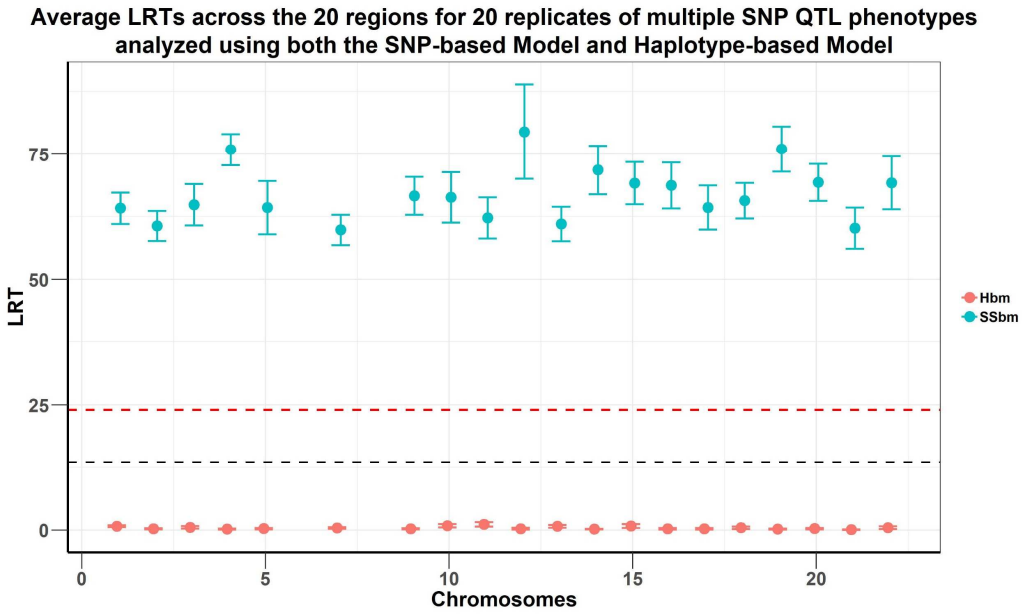
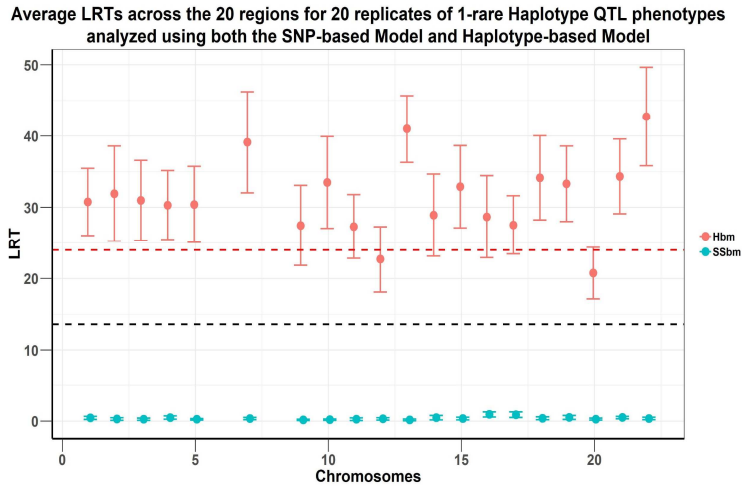
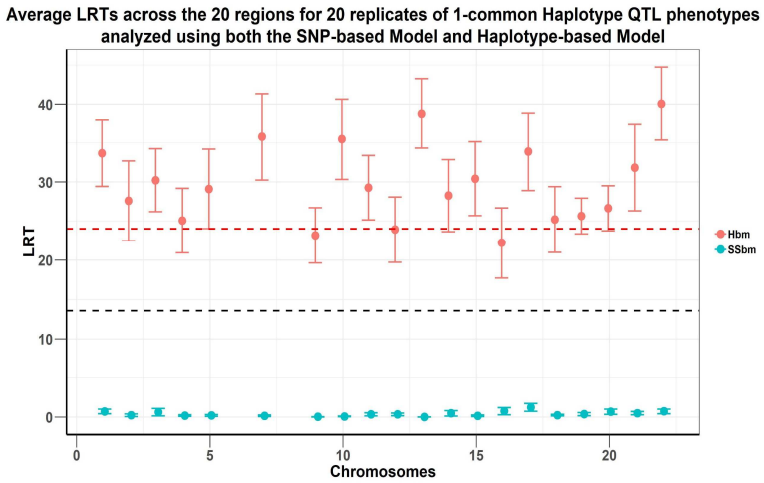



Figure 5.3a. Plots of average LRT statistics over replicates of QTL loci across the chromosomes for the 20 simulations of each of the two SNP QTL phenotypes. The red dashed lines are genome-wide significance at alpha of 0.05 and the black dashed lines are Bonferroni significance threshold (for 220 regions). The upper plot (i) is the 1-SNP QTL phenotype and the lower plot (ii) is the multiple SNP QTL phenotype. The two phenotypes are analysed using both the SNP based model (SSbm) (blue line) and the Haplotype based model (Hbm) (red line). The Haplotype based model fails to capture the simulated effects for the SNP QTLs.

i.




ii.




iii.

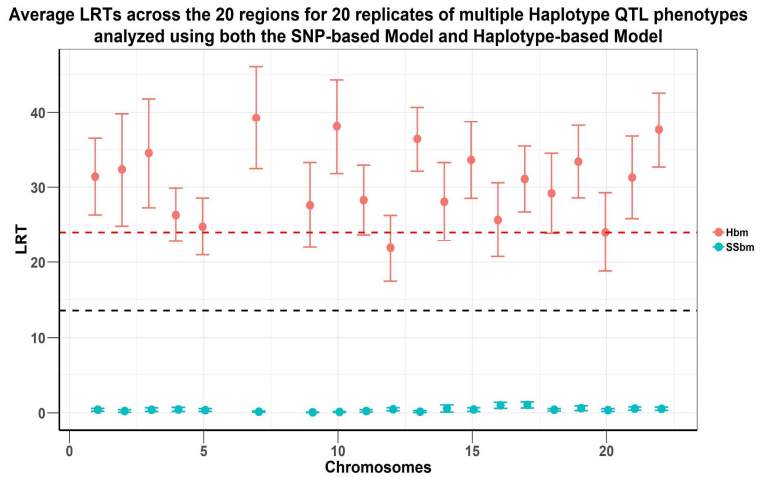
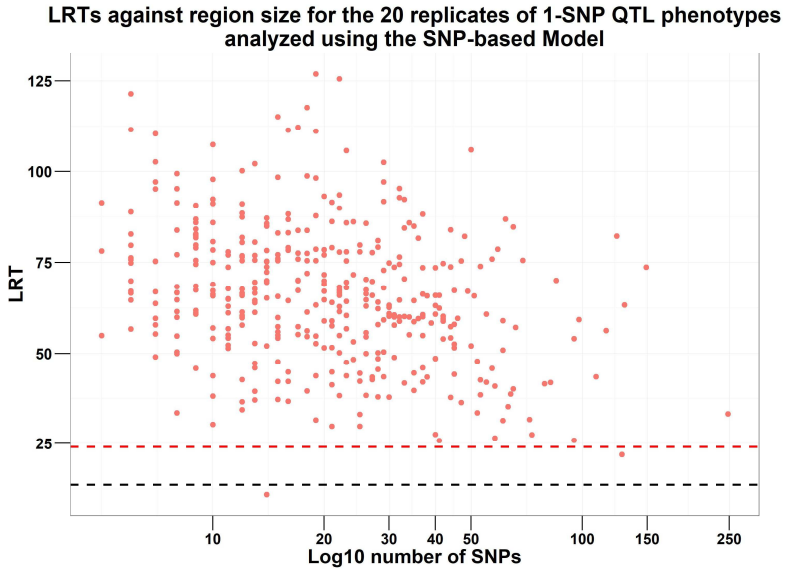



Figure 5.3b. Plots of average LRT statistics over replicates of QTL loci across the chromosomes for the 20 simulations of each of the three haplotype QTL phenotypes. The red dashed lines are genome-wide significance at alpha of 0.05 and the black dashed lines are Bonferroni significance threshold (for 220 regions). The plot (i) is the 1-rare haplotype QTL phenotype, the plot (ii) is the 1-common haplotype QTL phenotype and the plot (iii) is the multiple haplotype QTL phenotype. The three phenotypes are analysed using both the SNP-based model (blue line) and the Haplotype based model (red line). The SNP-based model fails to capture the simulated effects for the haplotype QTLs.

i.




ii.

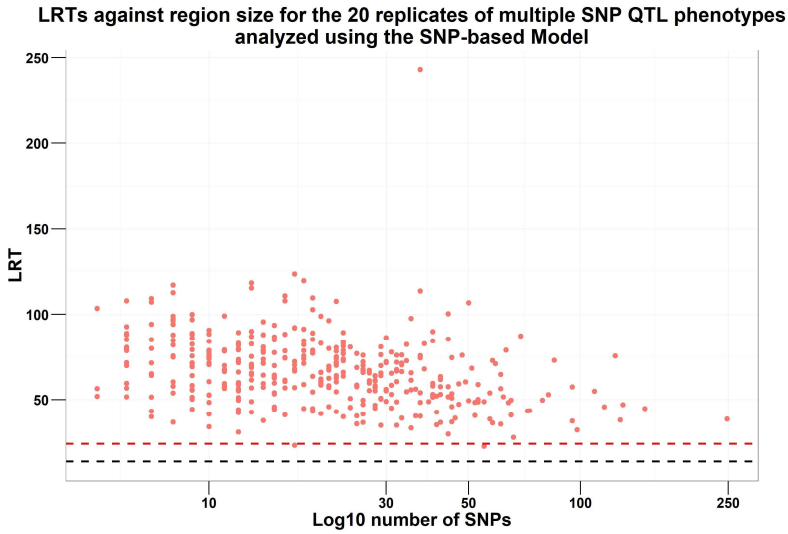
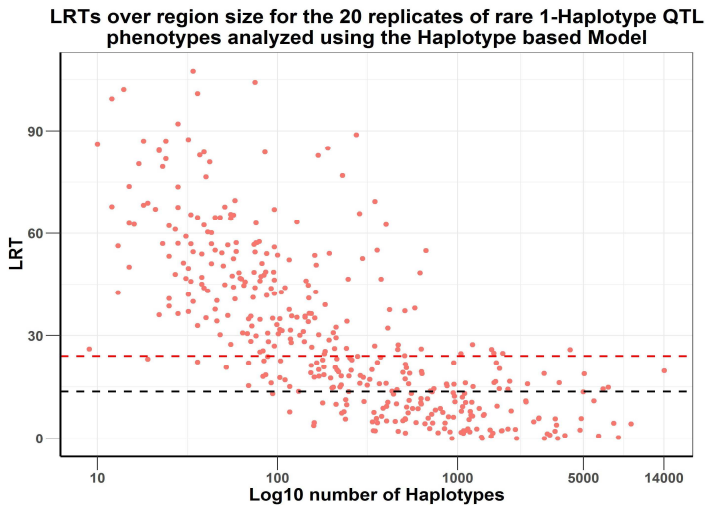
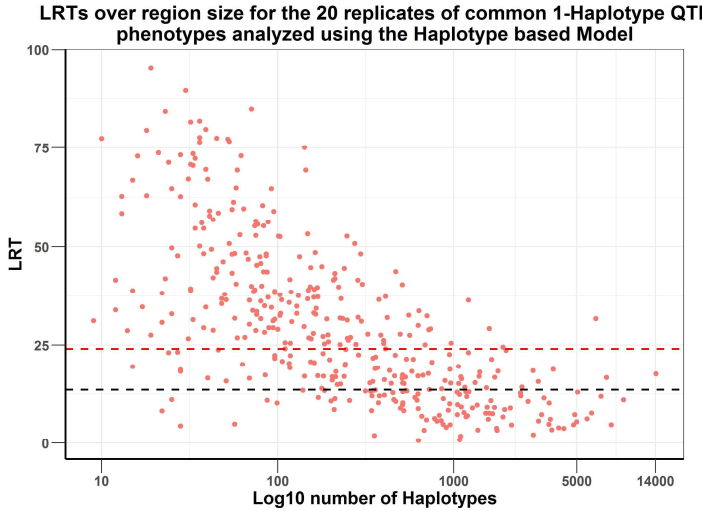



Figure 5.4a. Plots of LRT statistics against QTL region size for the 20 simulations (not averaged) of each of the two SNP QTL phenotypes. The red dashed lines are genome-wide significance at alpha of 0.05 and the black dashed lines are Bonferroni significance threshold (for 220 regions). Plot (i) is the 1-SNP QTL phenotype and the lower plot (ii) is the multiple SNP QTL phenotype. In both phenotypes the LRT statistic reduced with increasing region size.

i.




ii.




iii.

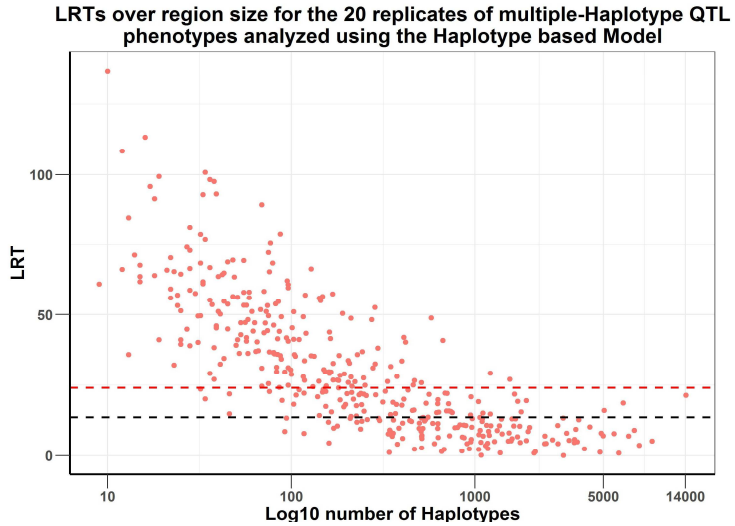



Figure 5.4b. Plots of LRT statistic against QTL region size for the 20 simulations of each of the three haplotype QTL phenotypes. The red dashed lines are genome-wide significance at alpha of 0.05 and the black dashed lines are Bonferroni significance threshold (for 220 regions). The plot (i) is the 1-rare haplotype QTL phenotype, the plot (ii) is the 1-common haplotype QTL phenotype and the plot (iii) is the multiple haplotype QTL phenotype. For all the three phenotypes the LRT statistic reduced with increasing region size.

For the SNP QTL phenotypes, the variance estimated by the Sbm in regions where I simulated effects were significantly different from zero (except two regions in both phenotypes) (Figure 5.5a). The variance was however overestimated for some regions, although there is no apparent relationship between the LRTs and the estimated variance (that is the estimated variance does not get closer to the simulated as the LRT increases).

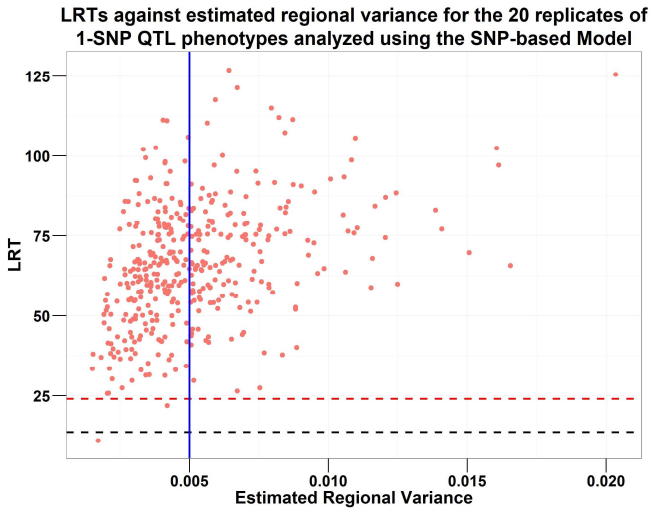
The regional variance estimates for the haplotype QTL phenotypes, on the other hand, improved with increasing LRTs (Figure 5.5b). The plots show that once the LRT passes the Bonferroni-corrected threshold for genome-wide significance (red horizontal dashed lines), the variance estimates approached the simulated value of 0.005 (blue vertical line).

Plots shown in Figure 5.6a show no relationship between region size (number of markers in the region) and the estimated regional variance for the SNP QTL phenotypes. The estimated regional variances for the haplotype QTL phenotypes, however, are inflated when region size has more than 5,000 different haplotypes, Figure 5.6b.

I further investigated whether the LRT statistics and the estimated regional variances are influenced by the allele frequencies of the QTLs. I show in Figure 5.7 that there is no relationship between allele frequencies and LRTs. The haplotypes simulated to have an effect in regions with an overestimated regional variance are marked with different colours on these plots. These haplotypes are shown to have rare haplotype frequencies and relatively low LRTs (Figure 5.7i and ii). Figure 5.7

shows that the QTLs for the 1-SNP QTL phenotypes and 1 common haplotype QTL phenotypes are well distributed across the MAF spectrum, whereas the QTLs for the 1 rare haplotype QTL phenotypes have a MAF distribution that is skewed towards zero.

i.




ii.

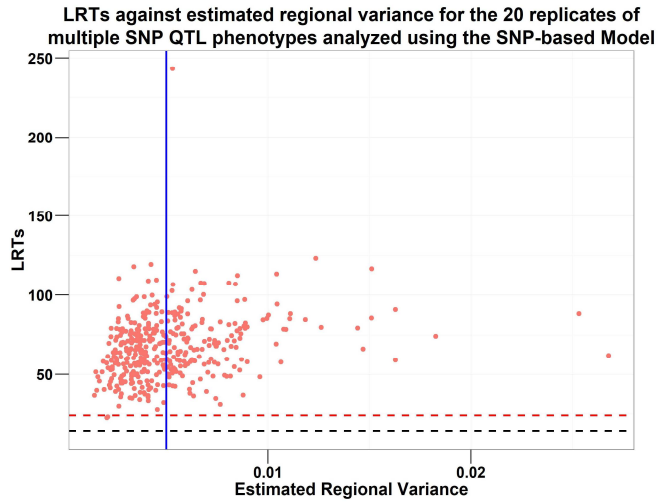
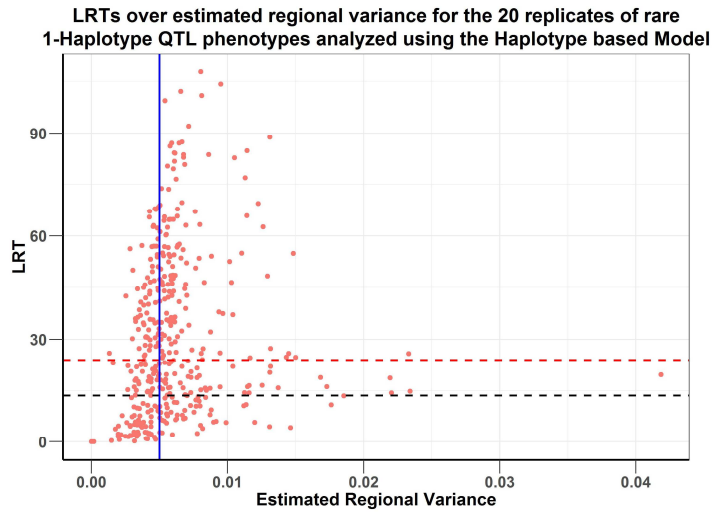
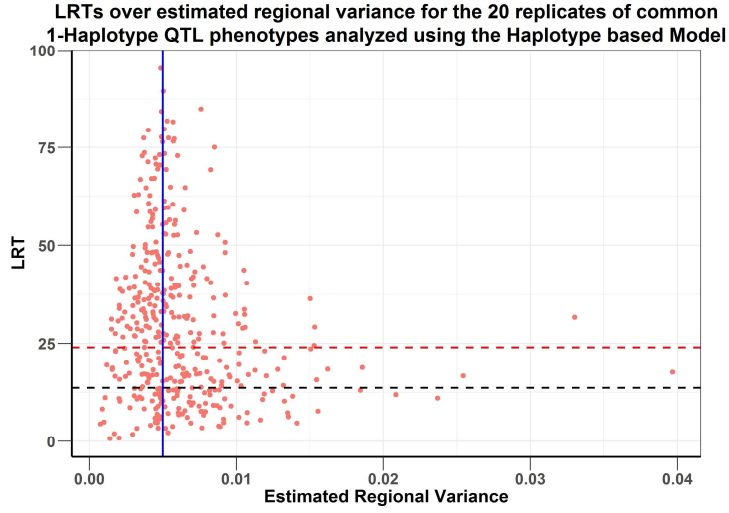



Figure 5.5a. Plots of LRT statistic against estimated regional variance for the 20 simulations of the single SNP QTL phenotype. The red dashed lines are genome-wide significance at alpha of 0.05, the black dashed lines are Bonferroni significance threshold (for 220 regions) and the blue vertical line is the simulated regional variance of 0.005. The plot (i) is the 1 – SNP QTL phenotype and the lower plot (ii) is the multiple SNP QTL phenotype. The estimated regional variance clustered closely around the simulated value for most of the regions. The variances of a few regions were overestimated in both phenotypes.

i.




ii.




iii.

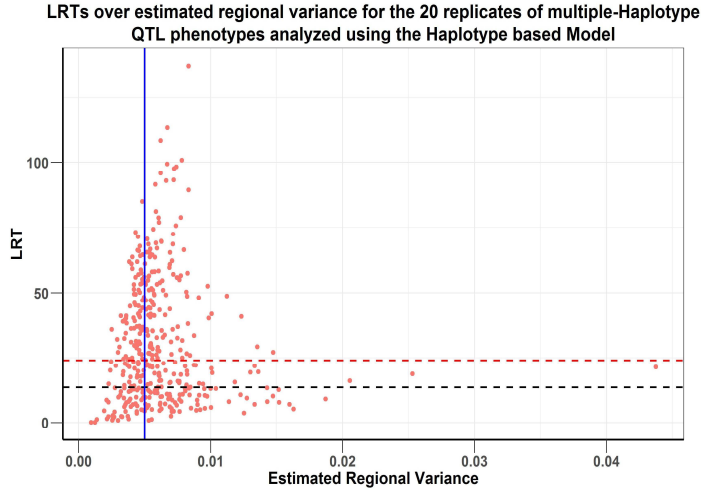
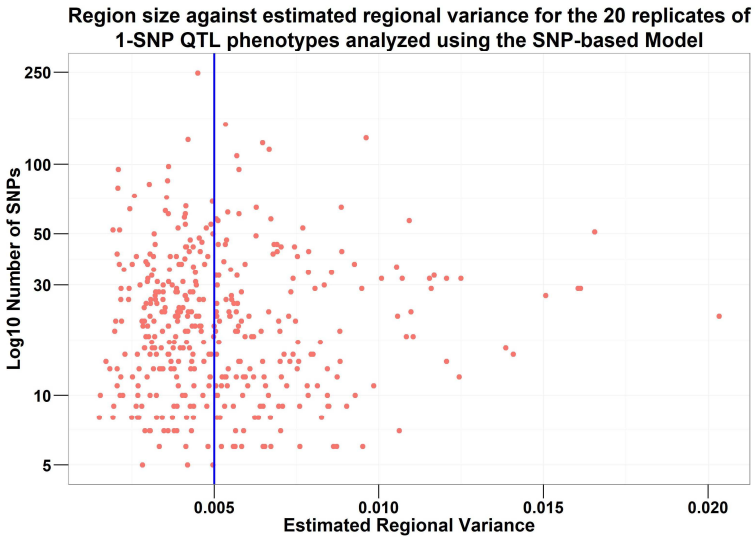



Figure 5.5b. Plots of LRT statistic against estimated regional variance for the 20 simulations of each of the three haplotype QTL phenotypes. The red dashed lines are for genome-wide significance at alpha of 0.05, the black dashed lines are Bonferroni significance threshold (for 220 regions) and the blue vertical line is the simulated regional variance of 0.005. The plot (i) is the 1-rare haplotype QTL phenotype, the plot (ii) is the 1-common haplotype QTL phenotype and the plot (iii) is the multiple haplotype QTL phenotype. The estimated regional variance clusters closely around the simulated value for the regions that pass the Bonferroni threshold.

i.




ii.

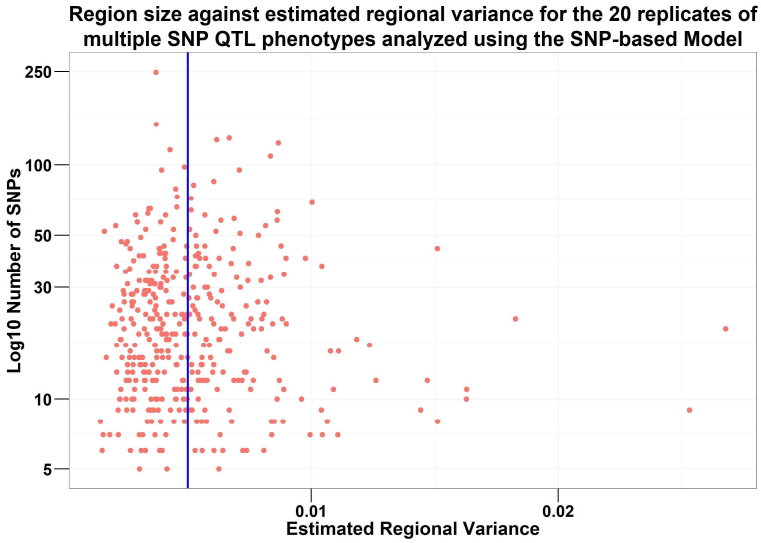
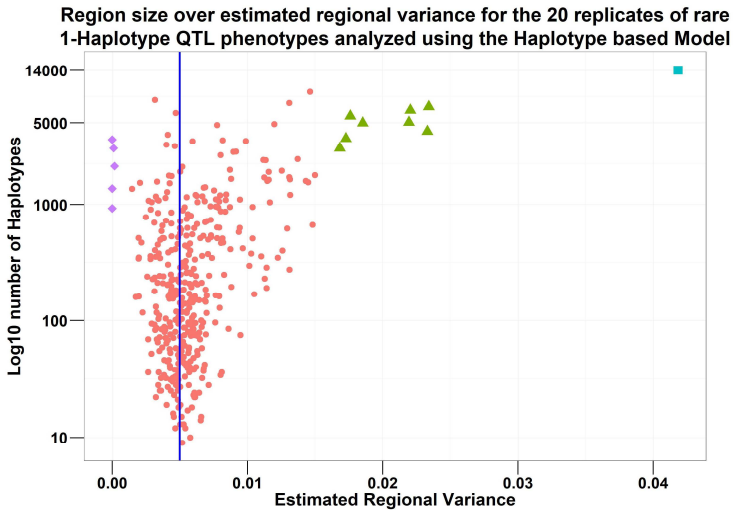
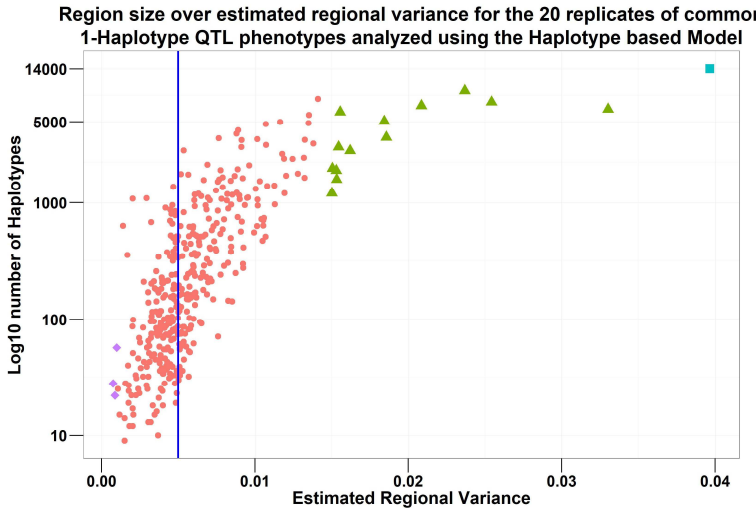



Figure 5.6a. Plots of region size against estimated regional variance for the 20 simulations of the two SNP QTL phenotype. The blue vertical line is the simulated regional variance of 0.005. The plot (i) is the 1-SNP QTL phenotype and the lower plot (ii) is the multiple SNP QTL phenotype. The two plots show there is no apparent relationship between estimated regional variance and region size.

i.




ii.




iii.

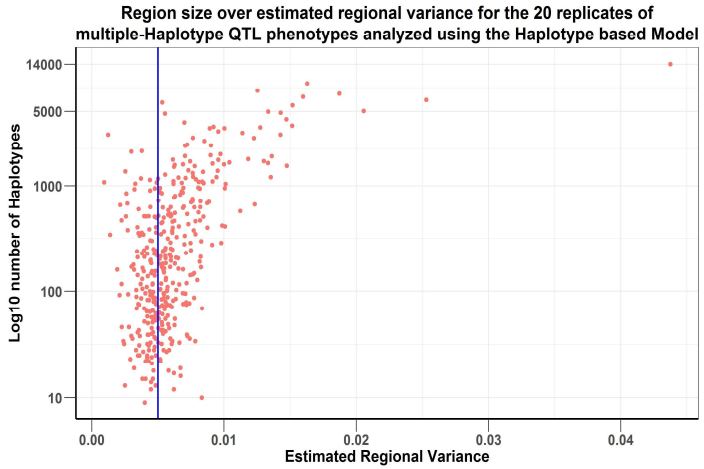
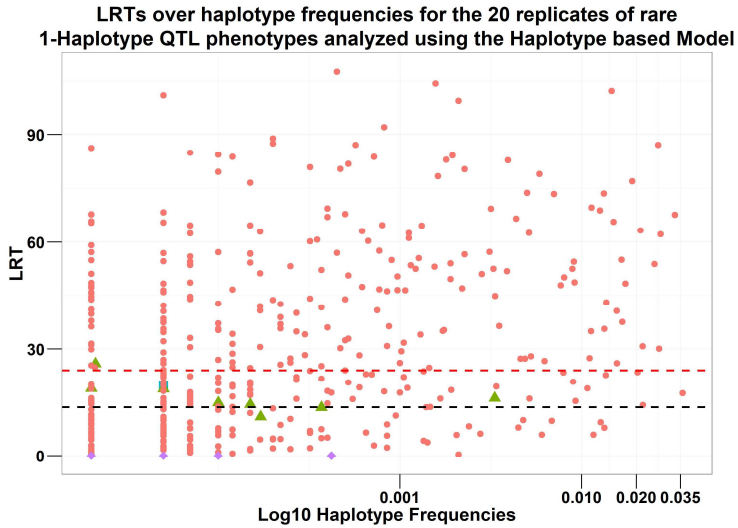
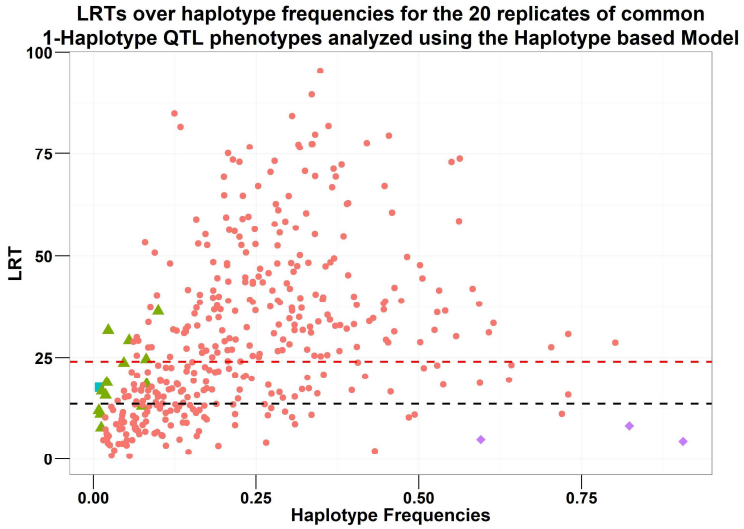



Figure 5.6b. Plots of region size against estimated regional variance for the 20 simulations of the three haplotype QTL phenotype. The blue vertical line is the simulated regional variance of 0.005. The plot (i) is the 1-rare haplotype QTL phenotype and the plot (ii) is the 1-common haplotype QTL phenotype. On these plots, the blue square point is the region with the largest overestimated variance, the green triangle points are regions with overestimated variance, red points are all other regions and purple points are regions with least variance estimates. The plot (iii) is the multiple haplotype QTL phenotype. The plots show that the estimates of the regional variances are likely to be inflated when the region size gets beyond 1000 haplotypes.

i.




ii.




iii.

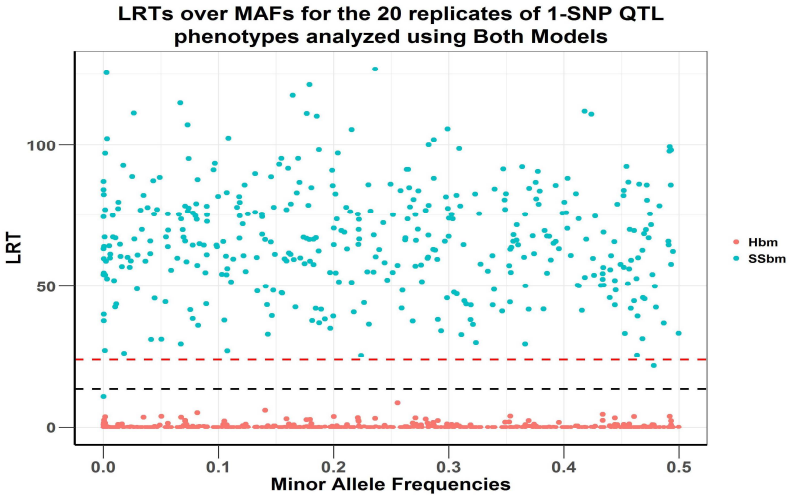



Figure 5.7. Plots of LRT statistic against QTL marker frequencies. The red dashed lines are genome-wide significance at alpha of 0.05 and the black dashed lines are Bonferroni significance threshold (for 220 regions). The plots (i) and (ii) are 1 rare and 1 common haplotype QTL phenotypes respectively. On these plots, the blue square point is the region with the largest overestimated variance, the green triangle points are regions with overestimated variance, red points are all other regions and purple points are regions with least variance estimates. The plot (iii) is the 1-SNP QTL phenotype analysed using the SNP based model (blue points) and the haplotype-based model (red points). The three plots show that there is no relationship between QTL marker frequencies and the LRT statistic.

When the region sizes used in the haplotype-based model were reduced in an analysis that restricted the natural haplotype block sizes to 20 or fewer SNPs per haplotype block, the haplotype-based model underestimated the regional variance at larger regions but did not offer any discernible improvement in the LRT statistics (Figure 5.8).

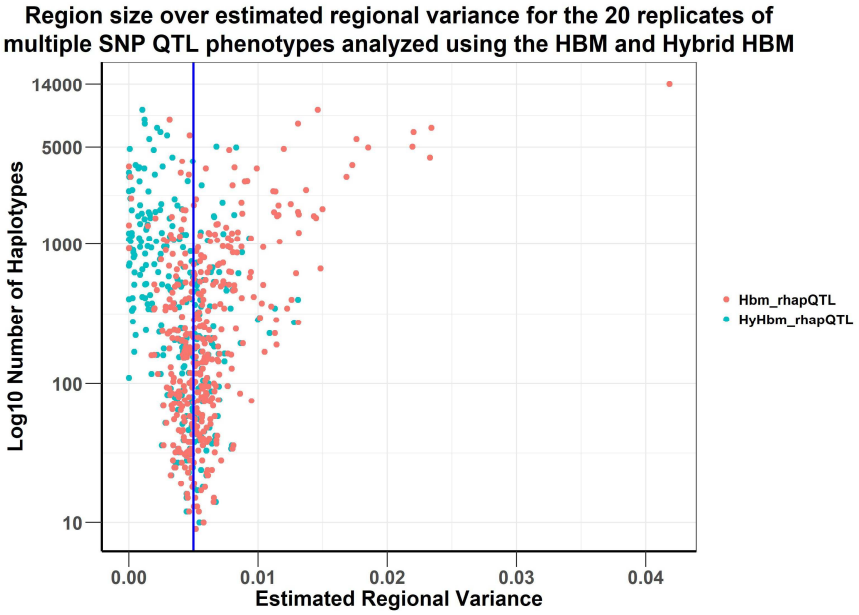
## 5.4 Discussion

I have proposed and implemented a genome-wide analytical method that analyses blocks of genomic regions using a regional GREML model (Cebamanos et al., 2014). The uniqueness in this method is that, genomic regions in my data are defined naturally by recombination hotspots drawn from a reference human genome and, I fit a regional GREML model that fits a haplotype-based GRM (Hbm) (Shirali et al., 2018). I also fit another regional GREML model that fits a SNP-based GRM (Sbm) to draw comparisons with the haplotype method.

I hypothesised that the haplotype-based model will complement conventional GWAS methods which are predominantly SNP-based. I investigated this hypothesis in a simulation study in which I simulated 20 replicates each of two types of SNP QTL phenotypes and three types of haplotype QTL phenotypes.

The results show that the two models can capture the effects of causal variants within genomic loci that are associated with the phenotype analysed. The usefulness of the SNP-based GREML model in analysing real phenotypes had been demonstrated and widely implemented in GWA research (Canela-Xandri et al., 2015; Cebamanos et al., 2014; Yang et al., 2010, 2011).

i.




ii.

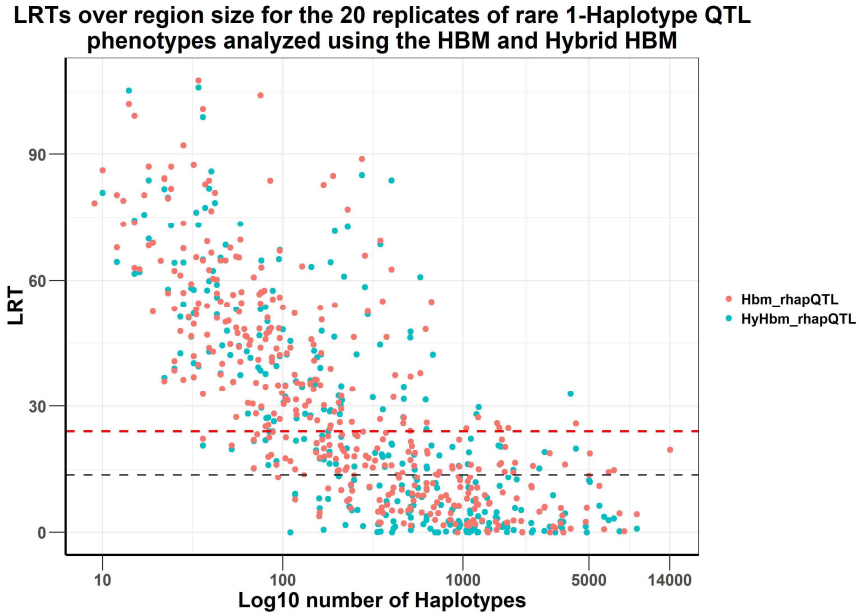
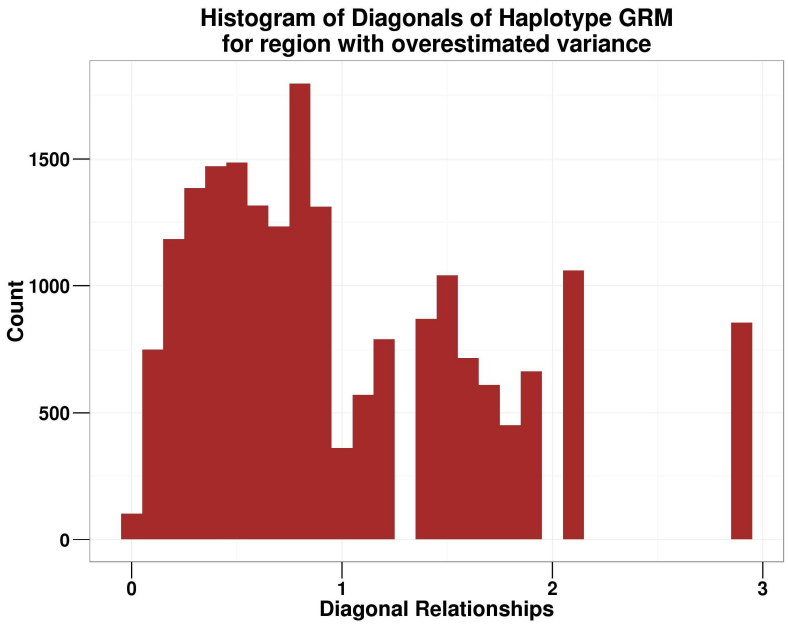



Figure 5.8. Plots for the 1-rare haplotype QTL phenotype analysed using the haplotype-based model (red points) and a hybrid variant of the haplotype-based model (blue points). The hybrid model broke larger regions (regions with more than 20 SNPs) into smaller regions of 20 or less SNPs and used that to determine the haplotypes. Each blue point represents the estimated variances or LRT for best sub-window within the bigger window. (i) is a plot of region size over estimated regional variance for the 20 simulations of the phenotype and (ii) is the plot of LRT statistic over QTL region size for the 20 simulations of each of the phenotype. The plot (i) shows that the hybrid model underestimates the regional variance at larger regions. The LRT statistics however do not improve very much over the default haplotype-based model.

i.




ii.

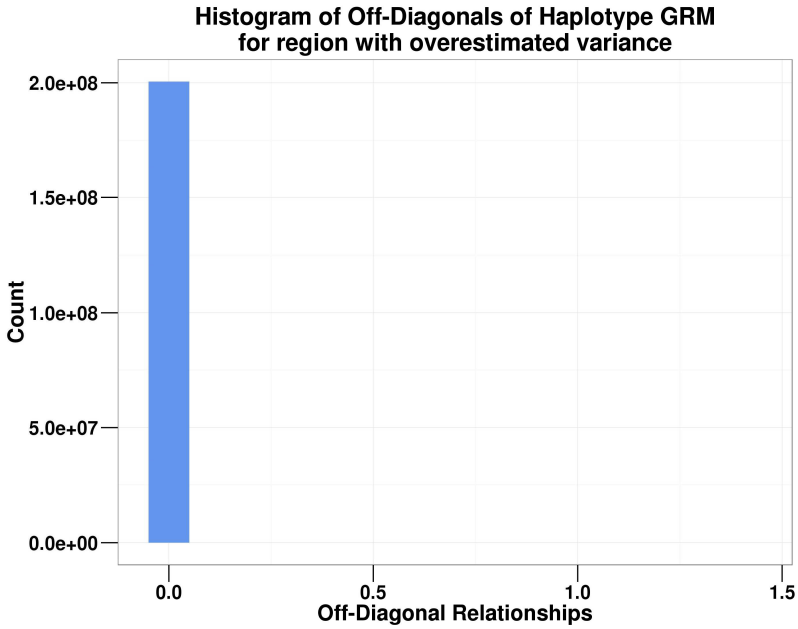



Figure 5.9. Histogram of counts of pairwise relationships for haplotype GRM for region with overestimated variance. (i) the diagonal relationships indicating relationship between individuals and themselves. (ii) the off-diagonal relationships indicating the relationship between an individual and everybody else. The lower plot shows there are lots zero kinships among individuals in the largest haplotype block.

The usefulness of the more novel haplotype-based GREML model (Shirali et al., 2018) for the analysis of real data is of particular interest because the results from this simulation study show that this model is capable of capturing causal loci with statistical significance which is very reassuring in terms of applying the model to real data and finding real effects.

The results also show that the two models are very specific to the type of marker effect they can capture. Figure 5.3a and b show that the haplotype-based model can specifically target haplotype effects which are mostly missed by SNP-based analysis and vice versa. These results, therefore, support my hypothesis that haplotype-based GREML models will complement SNP-based GREML models.

The GREML models use the covariance of genomic values calculated from SNPs that are identity by state (IBS) between pairs of study individuals. This IBS matrix is used to estimate the fraction of the phenotype variance explained by genotyped genomic markers (mostly SNPs) (Yang et al., 2010). The way haplotypes are defined in my setting makes it possible that in the same region, pairs of individuals may be related in the SNP-based GRM but not related in the haplotype-based GRM. This can generate major disparity between a SNP-based GRM and haplotype-based GRM within the same region in terms of which cluster of individuals are genetically similar and how different these cluster of individuals are from the rest of the population. The GREML model works by projecting the marker-derived genetic relationships between individuals onto their phenotypic differences to estimate the genetic variance (Yang et al., 2010). Thus, mismatching the simulated phenotype and the model will mean the relationship matrix won't reflect well the phenotypic differences

between the individuals. One would then observe very low or zero genetic variance estimates and highly insignificant LRT statistics as observed in the simulation study. This will also be expected when using real phenotypes.

The results shown in Figure 5.4a and b show that the LRT statistic decayed with increasing region size for both analytical models (even when phenotype simulation matches the analytical model). The rate of decay, however, was relatively higher in the haplotype-based analysis model. This observation may be explained by the structure of the GRMs. The pairwise genetic relationships provide the information for estimating the genetic variance and therefore there must be some genetic similarity between a subset of individuals in a region for the model to estimate the genetic variance within that region with statistical significance. If many pairs of individuals have “zero” (i.e. very small) kinships in a region then the genetic variance within that region cannot be estimated relatively with much statistical significance. The results showed the number of markers within regions increased by one or two orders of magnitude in the haplotype-based model compared to the SNP-based model. In large regions (with many markers) there are lots of very rare markers in the haplotype-based analysis. Therefore, since there won’t be adequate sharing of the rare haplotypes in these very large regions between individuals, there won’t be a generation of higher kinships in the haplotype-based GRMs. For instance, if you have a pair of individuals sharing one rare haplotype, then they would have a very high pairwise relationship estimate and their relationship with other individuals will be very small as well if there is no haplotype sharing. In such a GRM, there will be lots of near-zero pairwise relationships (Figure 5.9) and clusters of high pairwise

relationships which will affect the estimate of the variance and thus impact the LRT statistics.

The regional variances in the longer haplotype blocks were overestimated by the haplotype-based analysis model. These regions are known to potentially harbour lots of rare haplotypes, which could mean the haplotypes that are chosen as QTLs in the simulation are more likely to be rare. One could imagine that individuals who share a QTL haplotype would have greater genetic similarity and be unrelated to everyone else in the regional relationship matrix. The phenotype will be associated with this cluster of individuals and not associated with the rest and this should drive down the estimate of the variance and not overestimate it. However, because the long haplotypes blocks harbour lots of rare haplotypes, these haplotypes may be in long-range LD with variants in other regions (Sabeti et al., 2002) and thus explain some of the variance in other regions, whereas the shorter haplotypes will be common and won't be in long-range LD with variants in other regions. This view is in line with the neutral theory of molecular evolution (Kimura, 1983), where because it takes new variants such a long time to rise to high frequencies in the population, recombination would cause a considerable decay of the LD around them by the time they are common. Again, the GREML model normalises the marker genotypes and assumes that the effect size per normalised genotype follows a normal distribution. This indirectly assumes that a variant with lower MAF will have a larger allele effect and thus by design the model is likely to overestimate its effect. Although this model assumption is in harmony with a model of negative selection (Loewe, 2008) where selection acts on variants with larger effect sizes to keep them at lower frequencies,

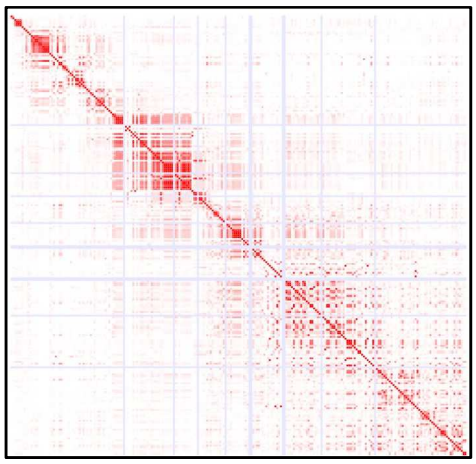
the effect sizes may be largely overestimated in some cases which may drive down the LRT statistic. I, therefore, observed very low test statistics for the long haplotype blocks (Figure 5.4b).

I attempted to mitigate this low test statistics by implementing a hybrid-Hbm that limited the haplotype window sizes to 20 SNPs or less. The results for this are shown in Figure 5.8 where the LRT for the sub-window explaining most of the variance within the bigger window is plotted. The hybrid model underestimated the variance for the size-limited haplotype windows and offered no improvement to the test statistics. The most obvious reason for this may be that breaking the large regions also breaks up the long haplotypes structures in such a manner that the resulting relationships estimated with the shorter haplotypes do not exactly match the ones obtained from long haplotypes used to simulate the phenotypes. For example, if you take a large haplotype window, the similarities between shared haplotypes may be driven by a combination of alleles at several SNPs along the haplotype. Therefore, breaking the window up such that a lot of individuals do not get the combination of the alleles in that haplotype driving the variance, makes detecting its effect difficult. I must mention here that, for any large window there were at least two sub-windows whose variance sum up approximately to give the total variance simulated. I reported the sub-windows with the highest test statistic because the sub-windows are most likely to be correlated and thus the sum of the variance and the covariance of all the sub-windows will end up overestimating the variance.

This may be an obvious problem for the Hbm, but this is not sufficient to downplay the usefulness of this analytical method in a real phenotype setting. In the real world, one could imagine most haplotype analyses going wrong because the haplotypes are not actual haplotypes; instead, they are imputed from genotype data. But this does not apply here because in this simulation study the imputed haplotypes are used for both simulation and analysis. Granted, it may be difficult to account for any uncertainty that arises during phasing when assessing the overall significance of any finding from this method, but when LD between markers is high the level of uncertainty is quite low (Balding, 2006). Therefore, for the large haplotype windows, one way to deal with the problem of low LRT is to use a recombination frequency of 5cM/Mb for these windows instead of the usual 10cM/Mb. This lower threshold generates windows with relatively high LD (Figure 5.10) and shorter haplotype lengths which can improve the test statistics.

In conclusion, I have implemented a regional GREML analysis that analyses regions in the genome delimited by natural recombination boundaries and shown that haplotype-based methods (Hbm) can capture portions of the genetic variance that may be missed by conventional SNP-based analysis when the simulated effect is not SNP. The Hbm, however, struggles to accurately capture causal effects at regions with very long haplotypes (haplotypes consisting of more than 20 SNPs) and I suggest utilising lower recombination threshold in such regions to alleviate this issue. The results from this simulation study show the usefulness of these models and I, therefore, implement these methods in the next chapter to analyse real phenotype data from GS: SFHS and UK Biobank data.

i.



ii.

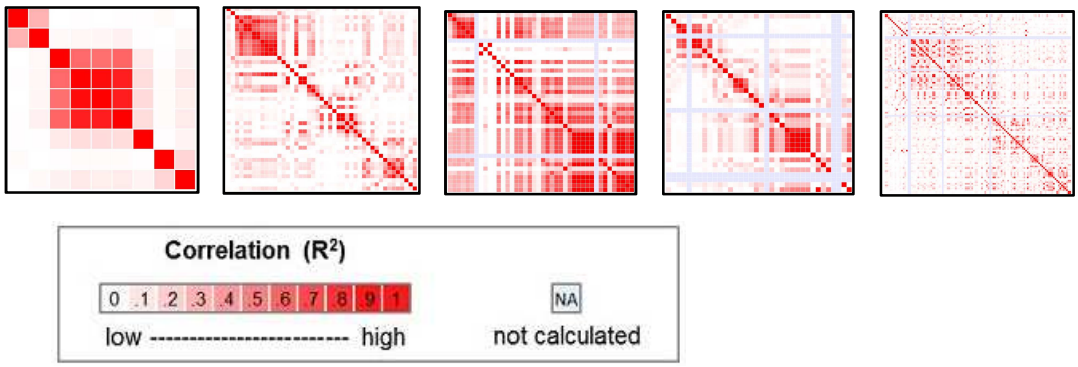
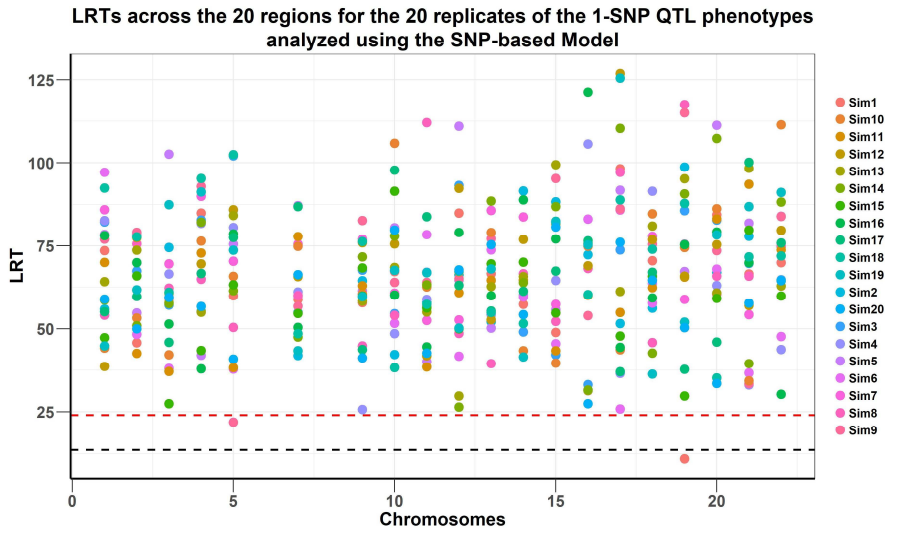



Figure 5.10. Plots of LD structure within largest haplotype window. (i) is the LD structure within a recombination threshold of 10cM/Mb. (ii) is the LD structure of the same region delimited by a recombination threshold of 5cM/Mb. The use of 5cM/Mb broke the window into 6 sub-windows and shown here is 5 plots (the third sub-window had just one SNP, so no plot is shown for it). From the plots it obvious that the LD structure large window improves when the recombination threshold is lowered.

i.




ii.

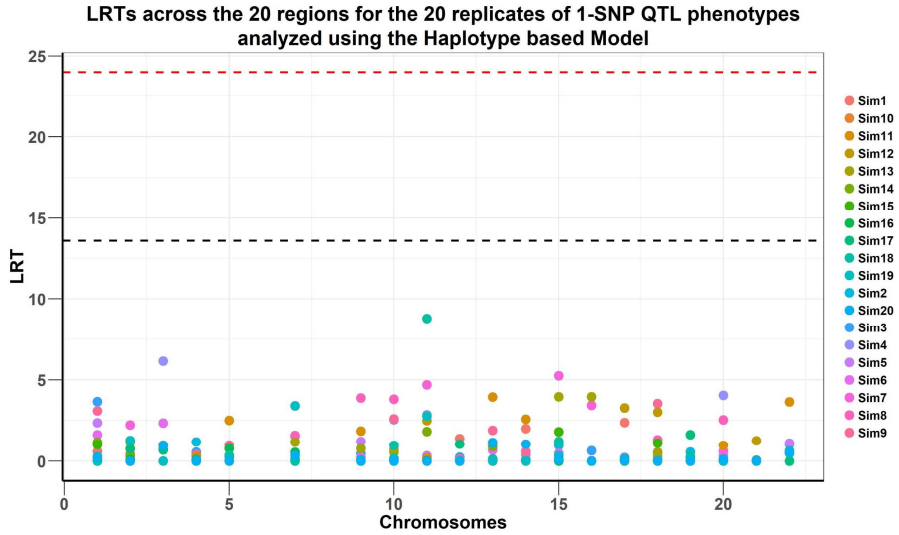
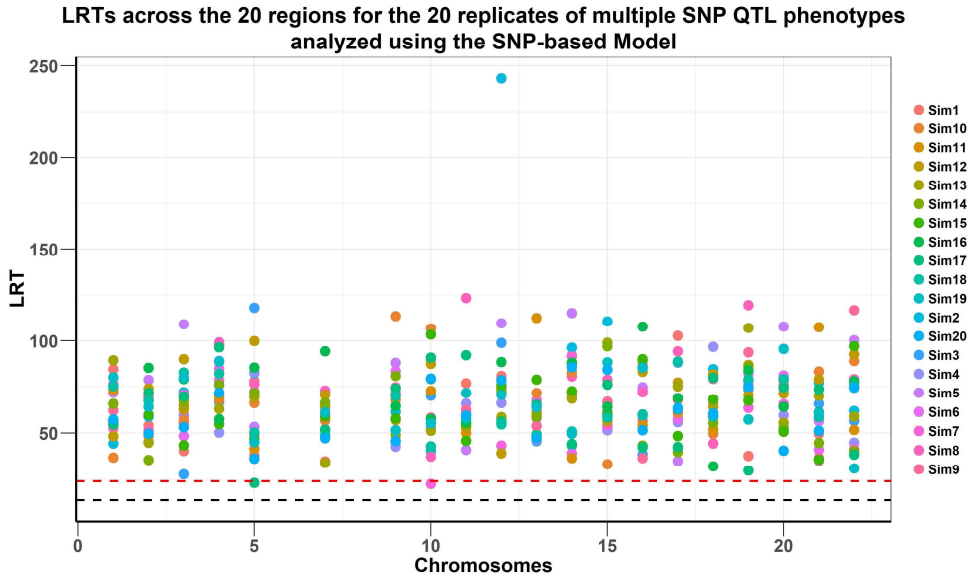



Figure 5.11a. Plots of Likelihood ratio test (LRT) statistic of QTL loci across the chromosomes for the 20 simulations of 1-SNP QTL phenotypes. The upper plot (i) is analysis done using the SNP-based model and the lower plot (ii) is analysis done using the Haplotype based model. The Haplotype based model fails to capture the simulated effects for the SNP QTLs.

i.




ii.

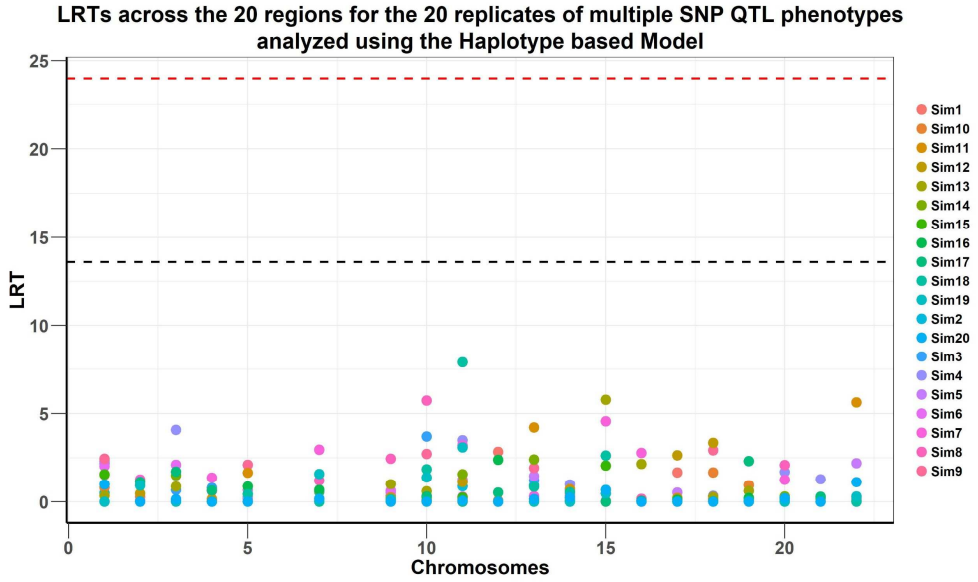
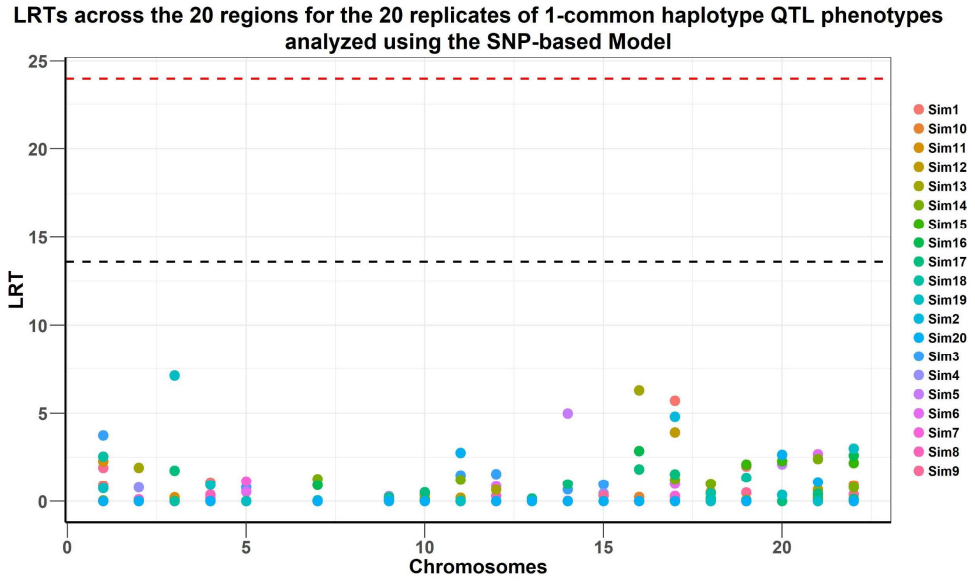



Figure 5.11b. Plots of Likelihood ratio test (LRT) statistic of QTL loci across the chromosomes for the 20 simulations of multiple-SNP QTL phenotypes. The top plot (i) is analysis done using the SNP-based model and the lower plot (ii) is analysis done using the Haplotype based model. Here again the haplotype-based model fails to capture the QTL effects simulated for multiple SNP QTL phenotypes.

i.




ii.

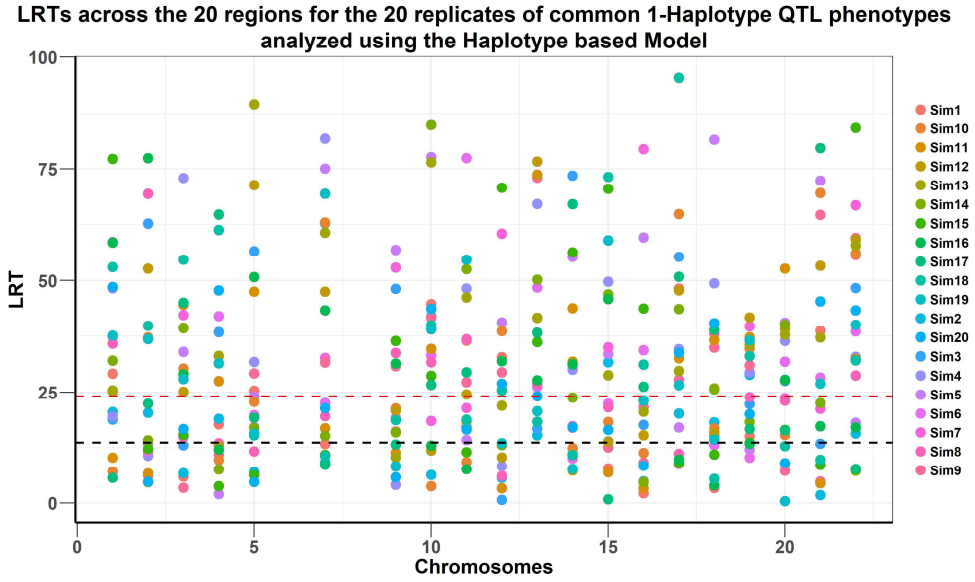
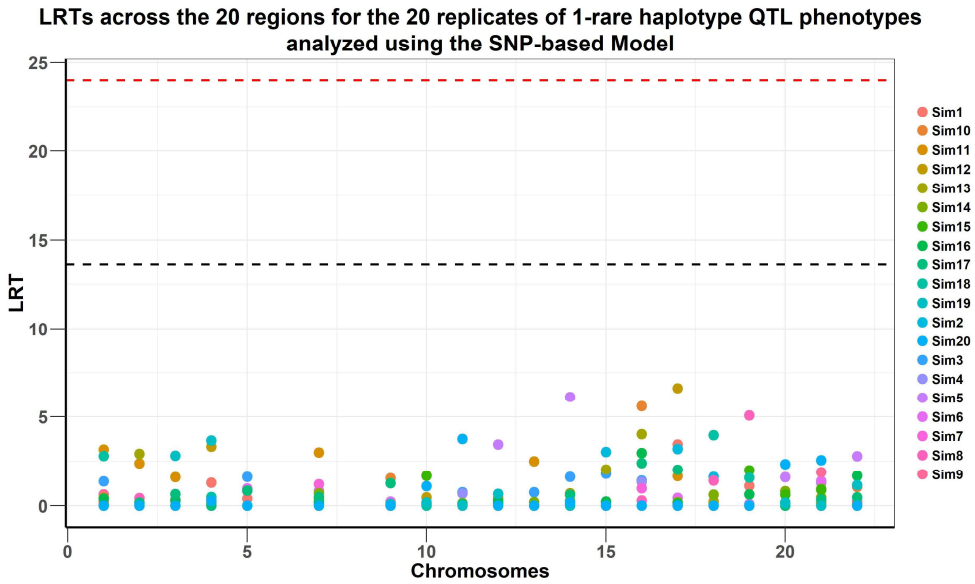



Figure 5.11c. Plots of Likelihood ratio test (LRT) statistic of QTL loci across the chromosomes for the 20 simulations of 1-common haplotype QTL phenotypes. The upper panel (i) is analysis done using the SNP-based model and the lower plot (ii) is analysis done using the Haplotype based model. In this phenotype the SNP-based model fails to capture the simulated QTL effects.

i.




ii.

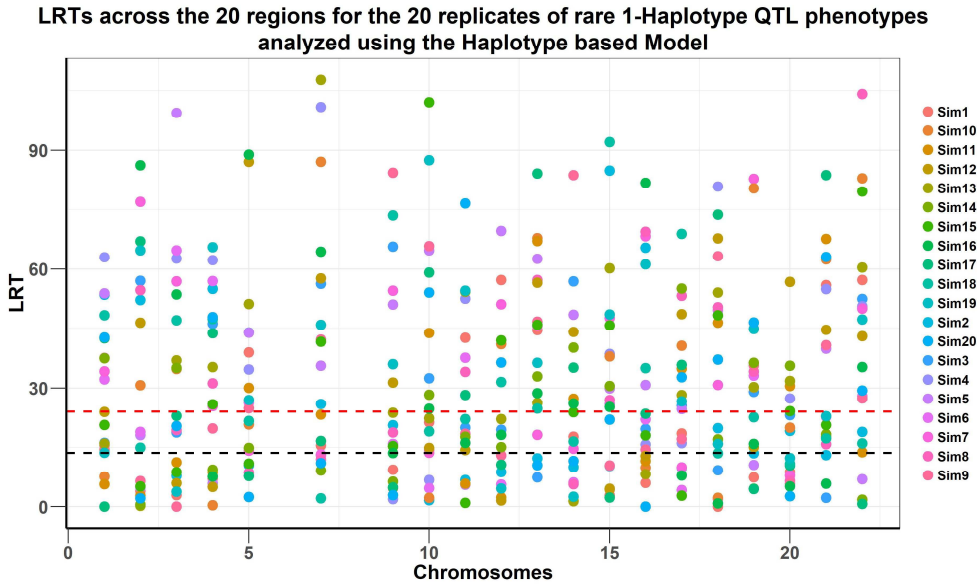
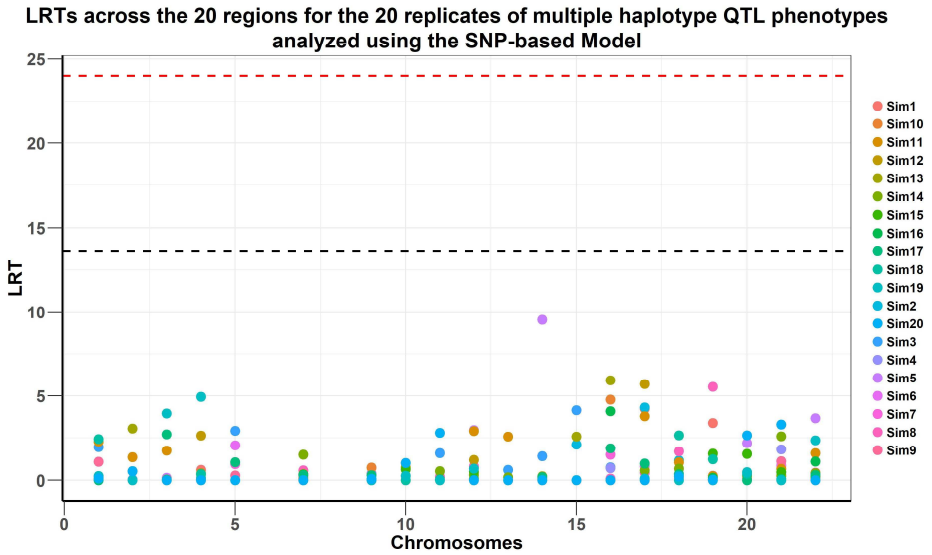



Figure 5.11d. Plots of Likelihood ratio test (LRT) statistic of QTL loci across the chromosomes for the 20 simulations of 1-rare haplotype QTL phenotypes. The upper panel (i) is analysis done using the SNP-based model and the lower plot (ii) is analysis done using the Haplotype based model. The haplotype-based model can capture the simulated QTL effect irrespective of the haplotype frequency.

i.




ii.

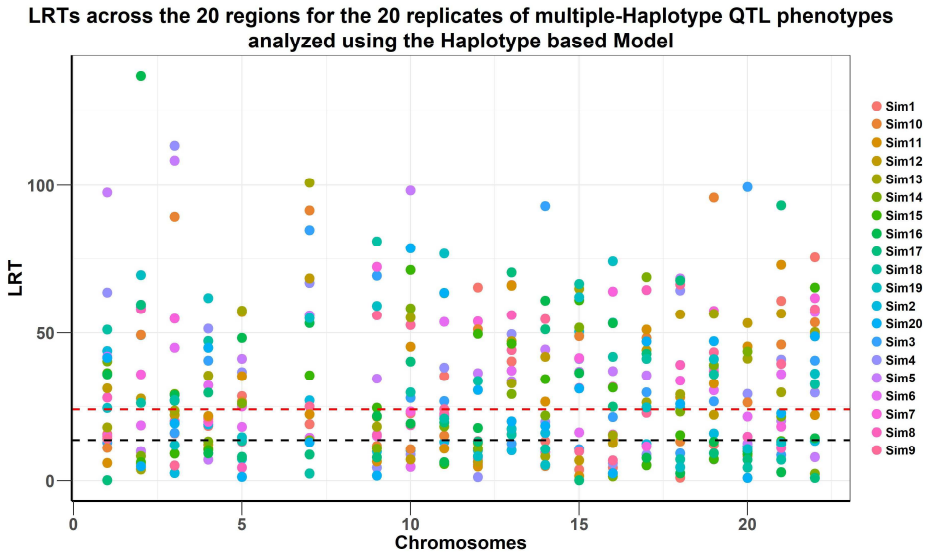



Figure 5.11e. Plots of Likelihood ratio test (LRT) statistic of QTL loci across the chromosomes for the 20 simulations of multiple-haplotype QTL phenotypes. The upper panel (i) is analysis done using the SNP-based model and the lower plot (ii) is analysis done using the Haplotype based model.

# Chapter 6

## 6 Regional heritability analysis of height and major depressive disorder phenotypes of GS: SFHS and UK Biobank using natural haplotype blocks defined by recombination boundaries

### 6.1 Introduction

GWA studies of height have been hugely successful with a lot of significant associations being found (Lin et al., 2017). That have offered some useful insights into the biology of height. The same cannot be said for major depressive disorder (MDD), however, for which uncovering genomic associations have proven difficult so far (Major Depressive Disorder Working Group of the Psychiatric GWAS Consortium et al., 2013).

The SNP heritability of MDD is estimated to be between 21% (Cross-Disorder Group of the Psychiatric Genomics Consortium et al., 2013) and 32% (Lubke et al., 2012), and this implies that there is some genetic contribution to the disease aetiology. The failure of GWAS to identify significant genomic loci for MDD has been put down to sample size and heterogeneity of the trait (Levinson et al., 2014). Large

sample size is needed before we can expect to identify multiple associations for MDD. Inferring from schizophrenia, Levinson et al. (2014) reasoned that an excess of 75,000 to 100,000 MDD cases is needed in a GWA study before we can expect to get some real answers.

More recent GWA studies of MDD have seen substantial sample size increases. A GWA study that analysed 322,580 UK Biobank individuals for three depression-related phenotypes found 14 independent loci significantly associated with MDD (Howard et al., 2018). Another study performed a GWA meta-analysis of 135,458 MDD cases and 344,901 controls and identified 44 independent loci significantly associated with MDD (Wray et al., 2018). These studies seem to provide support for the assertion that MDD GWA studies with sufficiently large sample sizes will detect multiple significant associations.

Increased sample size has offered some progress but that is not the only way forward for MDD GWA mapping. MDD is a very heterogeneous phenotype, and thus every MDD case will have a set of genetic and non-genetic risk factors exclusive to them (Levinson et al., 2014). The heterogeneity due to unique non-genetic risk factors reduces power to detect causal variants in GWA analysis across multiple cohorts (Wray and Maier, 2014). Also, the unique genetic risk factors will mean that at the population level, a lot of the genetic variants driving the disease will be rare. These potentially causal rare variants will be in weak linkage disequilibrium (LD) with genotyped SNPs and thus GWA analysis of SNPs will not have sufficient power to detect them. The answer to uncovering genetic risk loci for MDD, therefore, does not

lie in large sample size GWA studies alone, other analysis strategies to unearth MDD risk loci need to be developed.

Haplotypes are sets of linked SNP alleles on the same chromosome. For a trait like MDD which may be driven by lots of rare variants, seeking associations with haplotypes might be a better approach than seeking associations with SNPs in uncovering loci with effect. This is because although rare variants will be in weak LD with genotyped variants, they might be in strong LD with haplotypes. A genome-wide haplotype-based association analysis of MDD in 18,773 individuals identified two haplotypes to be significantly associated with MDD (Howard et al., 2017). There is, therefore, some promise in GWA study of MDD by employing haplotypes.

The usefulness of other mapping methods like the regional GREML analysis in uncovering significant loci in MDD has also been shown (Zeng et al., 2017a, 2017b), where gene loci significantly associated with MDD were identified.

This study, therefore, performs a regional GREML analysis of MDD and height in about 20,000 GS: SFHS individuals using natural haplotypes blocks defined by recombination hotspots in the genome. The aim is to be able to capture novel genetic variants that may be having an effect on these traits and improve the estimates of the genetic components of the variation in these traits. I also compare the results to the results of analyses of height and MDD done using a mixed linear model, GBLUP and BayesR. The study will then seek replication of any association results obtained in the GS: SFHS in the UK Biobank Leeds subpopulation cohort which is the largest sub-cohort of UK Biobank.

## 6.2 Methods

### 6.2.1 Study cohorts

#### *6.2.1.1 Genotyping, quality control and phasing of Generation Scotland: Scottish Family Health Study dataset*

The data from the Generation Scotland: Scottish Family Health Study comprised of 23,960 participants recruited from Scotland (Smith et al., 2006). The DNA from about 20,032 of the participants had been analysed using the Illumina HumanOmniExpressExome8v1-2_A chip (~700K genome-wide SNP chip) (Smith et al., 2012).

Quality control (QC) excluded SNPs and individuals with a call rate less than 98%, SNPs with minor allele frequency (MAF) less than 1% and SNPs that were out of Hardy-Weinberg equilibrium (p-value < 0.000001). A total of 555,091 autosomal SNPs passed quality control for downstream analysis. Phasing of the GS: SFHS data was done using SHAPEIT2 (Delaneau et al., 2013). Haplotype blocks were defined using recombination hotspots with a recombination rate of 10cM/Mb inferred from the Reference Consortium Human Build 37 (International Human Genome Sequencing Consortium, 2004). Haplotypes within blocks were determined using the phased data.

#### *6.2.1.2 Genotyping, quality control and phasing of UK Biobank dataset*

The full UK Biobank (Sudlow et al., 2015) dataset contains genotype information for about 488,377 participants. The genotypes of about 50,000 participants were assayed using the UK BiLEVE Axiom array (Wain et al., 2015) by Affymetrix (807,411 SNPs). The remaining participants in the cohort were genotyped using the UK Biobank Axiom array (825,927 SNPs) (Bycroft et al., 2017).

The cohort of participants from Leeds, the largest recruitment centre (in terms of number of participants) of the UK Biobank, were used as replication cohort to assess those genomic regions that were identified as associated for MDD within the GS: SFHS by the haplotype-based analysis model with p-value $< 5 \times 10^{-5}$. Pre-QC of the data excluded individuals identified as non-white British using genotypes rather than self-reported ancestry. It also removed outliers based on heterozygosity and missingness and sex mismatch between self and inferred sex. Individuals who had withdrawn consent were also excluded. A second QC removed individuals and SNPs with a call rate less than 98%, SNPs with a MAF less than 1% and SNPs that were out of Hardy-Weinberg equilibrium (p-value < 0.000001). A total of 38,010 individuals and 621,890 SNPs passed QC and were used for downstream analysis. The data had been phased using a SHAPEIT3 (Marchini, 2015).

### 6.2.2 Phenotype definition

MDD status for GS: SFHS participants was assigned following an initial mental health screening questionnaire with the questions: “Have you ever seen anybody for emotional or psychiatric problems?” or “Was there ever a time when you, or someone else, thought you should see someone because of the way you were feeling or acting?” Participants who answered yes to one or more of the screening questions were further interviewed by the structured clinical interview for the diagnosis of mood disorders (SCID) (First et al., 2002). A total of 18,725 participants (2,603 MDD cases and 16,122 controls) were retained for analysis for MDD.

A total of 19,944 participants from the GS: SFHS were analysed for height. The phenotypes were controlled for sex, age, age$^2$, and first 20 principal components calculated from the genomic covariance matrix of study participants.

A broad MDD phenotype was defined for UK Biobank participants in which depression cases were diagnosed for participants who answered yes to either of these questions: “Have you ever seen a GP for nerves, anxiety, tension or depression?” or “Have you ever seen a psychiatrist for nerves, anxiety, tension or depression?” A total of 13,723 cases and 24,287 controls were defined for the UK Biobank cohort.

### 6.2.3 Regional GREML analysis

Regional GREML analysis using fixed region sizes in the genome has been shown to be a good mapping method for finding local genetic effects (Nagamine et al., 2012). The model uses both local and genome-wide GRMs in the analyses to map genetic loci with effect. A regional GREML method that uses recombination hotspots to define natural region sizes in the genome has been successfully used in the analysis of MDD to map genetic loci (Zeng et al., 2017a, 2017b).

This study utilizes a regional GREML model that defines natural regions delimited by recombination hotspots to analyse height and MDD. Two types of regional GREML models are fitted in turn to the phenotypes. The first model uses SNPs to construct local genetic relationships between study individuals and the second model defines local relationships amongst individuals using haplotypes.

The significance of a region was tested with the likelihood ratio test (LRT). The model fits the effects of markers in each region as random and the effects of all SNPs in the background outside a region analysed also as random. The LRT tests for the significance of a random regional effect by comparing a model with both the regional and whole-genome effects fitted against a model in which only the whole-genome effect is fitted.

The p-values obtained from LRTs were used to generate genome-wide association plots for both models for each phenotype. The genome-wide significance threshold was calculated to be LRT = 23.88 (p-value $< 1.02 \times 10^{-6}$) for the SNP-based model and LRT = 22.54 (p-value $< 2.04 \times 10^{-6}$) for the haplotype-based model using a Bonferroni correction for testing 48,772 and 24,513 regions respectively (a lesser number of regions were tested for the Hbm because I skipped regions with one SNP). The suggestive significance of a region was set at an LRT = 16.5 (p-value $< 5 \times 10^{-5}$).

### 6.2.4 Mixed linear model, GBLUP and BayesR analysis of height and MDD

The height and MDD phenotypes from the GS: SFHS were analysed using the three GWAS models described in the previous chapter: the mixed linear model analysis, the genomic best linear unbiased predictor (GBLUP) and BayesR. The p-values of association and 1 minus the posterior inclusion probabilities (PIP) of the zero-effect distribution were used to generate genome-wide association plots for the two phenotypes. The genome-wide significance threshold was set at p-value $< 5 \times 10^{-8}$.

### 6.2.5 SNP-based association test of SNPs in most significant regions identified by the Haplotype-based model for MDD

A linear mixed effects model was used to test for phenotype association with the SNPs in the genome-wide significant regions identified by the haplotype-based regional GREML analysis. The effects of the regional SNPs and covariates such as sex, age, age$^2$ and the first 20 principal components calculated from the genomic covariance matrix of study participants were fitted as fixed, and the background polygenic effect was fitted as random in the linear mixed effects model association analysis using GCTA (Yang et al., 2011).

## 6.3 Results

### 6.3.1 Regional GREML analysis of GS: SFHS

The heritability estimates for height and MDD are 81.4% and 13.8% respectively (Table 6.1). The chromosome number and the heritability estimates for the genome-wide significant regions are shown in Table 6.1. For the two traits, there were no overlaps between regions identified as significant by both models. This reaffirms my hypothesis that the haplotype-based model is complementary to the SNP-based models in mapping out associated genomic loci.

The regional GREML results for height and MDD are presented as plots of minus-Log10 of LRT p-values (Figure 6.1 and Figure 6.2). The plots for the analysis using the two models, SNP-based and haplotype-based models are shown. The results for the mixed linear model analysis (MLMA) and BayesR are also shown in Figure 6.1 and Figure 6.2 for the two traits.

The results for height show that 14 regions passed the Bonferroni-corrected genome-wide significance threshold in the analysis using the SNP-based regional GREML model. No region was genome-wide significant for height when analysed with the haplotype-based model, but three regions were significant at the suggestive level. There are five genome-wide significant regions for height analysed with the MLMA. The BayesR analysis of height showed several regions with PIPs $> 0.5$, with about seven regions having PIPs $> 0.9$.

Table 6.1. The heritability estimates of traits under the two models. The columns are the type of heritability estimate, chromosome (Chr) number, heritability estimates and standard errors for the height and MDD under the two models.

| | | Height | | MDD | |
|---|---|---|---|---|---|
| **Heritability** | **Chr** | **Sbm** | **Hbm** | **Sbm** | **Hbm** |
| **Total $h^2$** | | 81.36 (0.9) | 81.35 (0.9) | 13.82 (1.3) | 13.81 (1.3) |
| **$h^2$ GW significant regions for MDD** | 3 | | | | 1.5 (0.3) |
| | 20 | | | 0.35 (0.26) | |
| **$h^2$ GW significant regions for height** | 6 | 0.16 (0.1) | | | |
| | 20 | 0.32 (0.2) | | | |
| | 3 | 0.19 (0.1) | | | |
| | 4 | 0.5(0.24) | | | |
| | 18 | 0.24 (0.12) | | | |
| | 2 | 0.26 (0.15) | | | |
| | 6 | 0.6 (0.28) | | | |
| | 6 | 0.43 (0.18) | | | |
| | 15 | 0.72 (0.43) | | | |
| | 13 | 0.14 (0.15) | | | |
| | 15 | 0.19 (0.1) | | | |
| | 6 | 0.29 (0.22) | | | |
| | 15 | 0.1 (0.1) | | | |
| | 6 | 0.37 (0.18) | | | |

From the plots in Figure 6.1, it can be seen that the association hits are more enriched in the SNP-based analyses than in the haplotype-based analysis for height. This can be explained by the fact that height is a trait driven by a lot of common genetic variants which may be in LD with genotyped SNPs on SNP chips

(disproportionately enriched for common SNPs) and thus SNP-based analyses should capture the effect of most of the genetic variants in height.

For MDD, one region passed the Bonferroni-corrected genome-wide significance threshold for the analysis done with the SNP-based regional GREML model. Again, one region passed the genome-wide significance threshold when MDD was analysed with the haplotype-based model, however, 20 regions were significant at the suggestive p-value. There are six regions that are significantly associated with MDD at the suggestive level when analysed with the MLMA. The most significant region for MDD was on chromosome 10 for the BayesR analysis. This same region was found to be the most significant for MDD for the MLMA. The SNP driving this association is rs1331328 with p-value $< 1.01 \times 10^{-6}$ for the MLMA and PIP $> 0.332$ for the BayesR analysis. The SNP was again found to be lying in a region that was nearly genome-wide significant at the suggestive level for (p-value $< 6.16 \times 10^{-5}$) MDD in the SNP-based GREML analysis. The SNP lies in the *CACNB2* gene which has been reported to be associated with five major psychiatric disorders including major depressive disorder (Cross-Disorder Group of the Psychiatric Genomics Consortium, 2013). The gene has also been reported as an emerging pharmacological target for mental disorders (Soldatov, 2015). The gene region was however not replicated in the haplotype-based GREML analysis.

The association hits for MDD are more enriched as one moves from the SNP-based analyses to the haplotype-based analysis. This can be down to the fact that the genetic variation in MDD may be driven by rare genetic variants which may be in LD with haplotypes and thus their effects are captured by haplotype-based models.

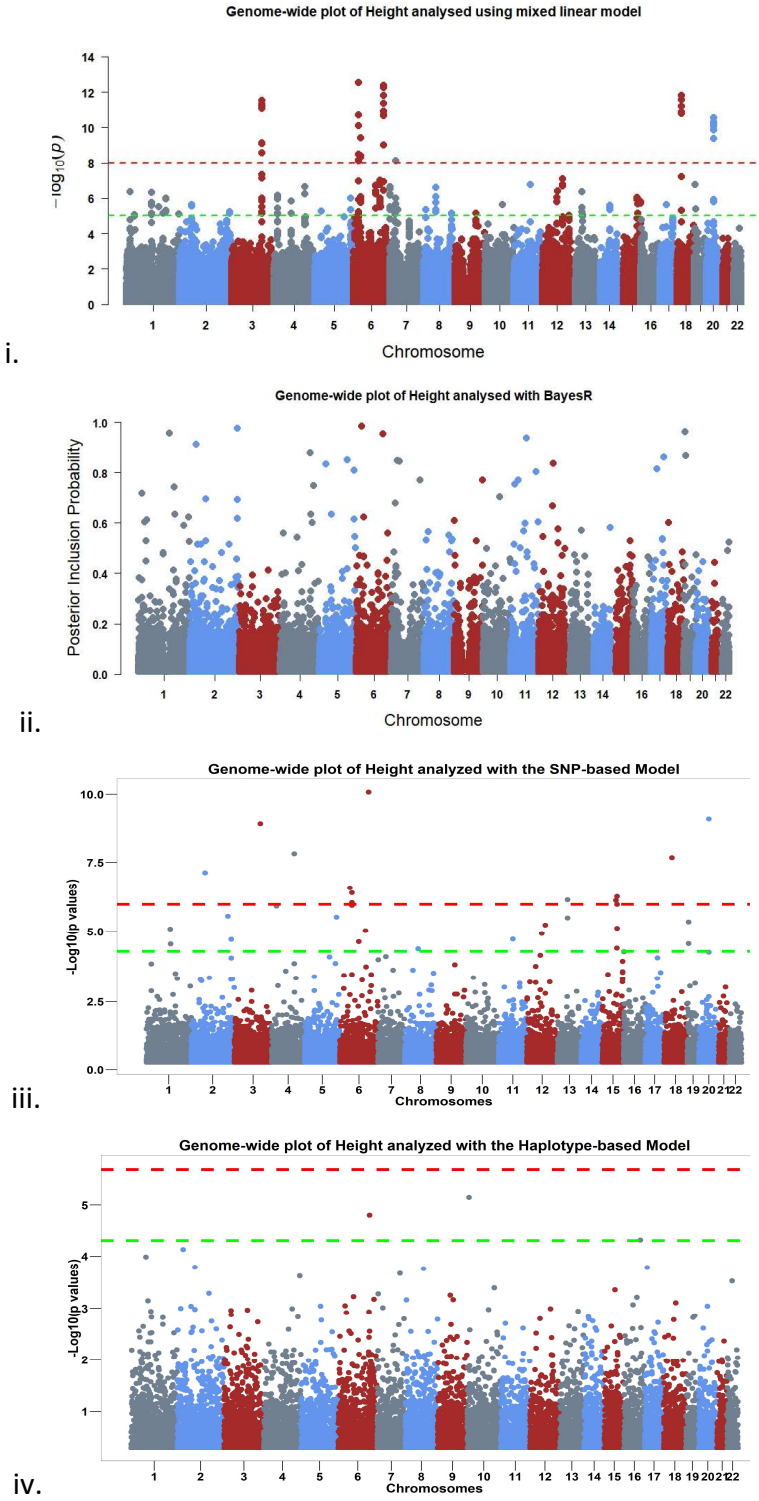



Figure 6.1. The genome-wide evidence of haplotype block association for height. Analysis done with i. mixed linear model (MLM) ii. BayesR iii. SNP-based regional GREML model and iv. Haplotype-based regional GREML model. The points on the BayesR plot are plots of the posterior inclusion probability of SNPs. The points on the remaining plots are plots of -log10 of the p-values of SNP tested for MLM analysis and regions tested for the regional GREML analyses. The red dashed lines are the Bonferroni-corrected genome-wide significance threshold and the green dashed lines are suggestive significance threshold calculated to be p-value $< 5 \times 10^{-5}$. The association hits are more enriched in the SNP-based analyses than in the haplotype-based analysis. This can be explained by the fact that height is a trait driven by a lot of common genetic variants which may be in LD with genotyped SNPs on SNP chips and thus SNP-based analyses should capture the effect of most of the genetic variants in height.

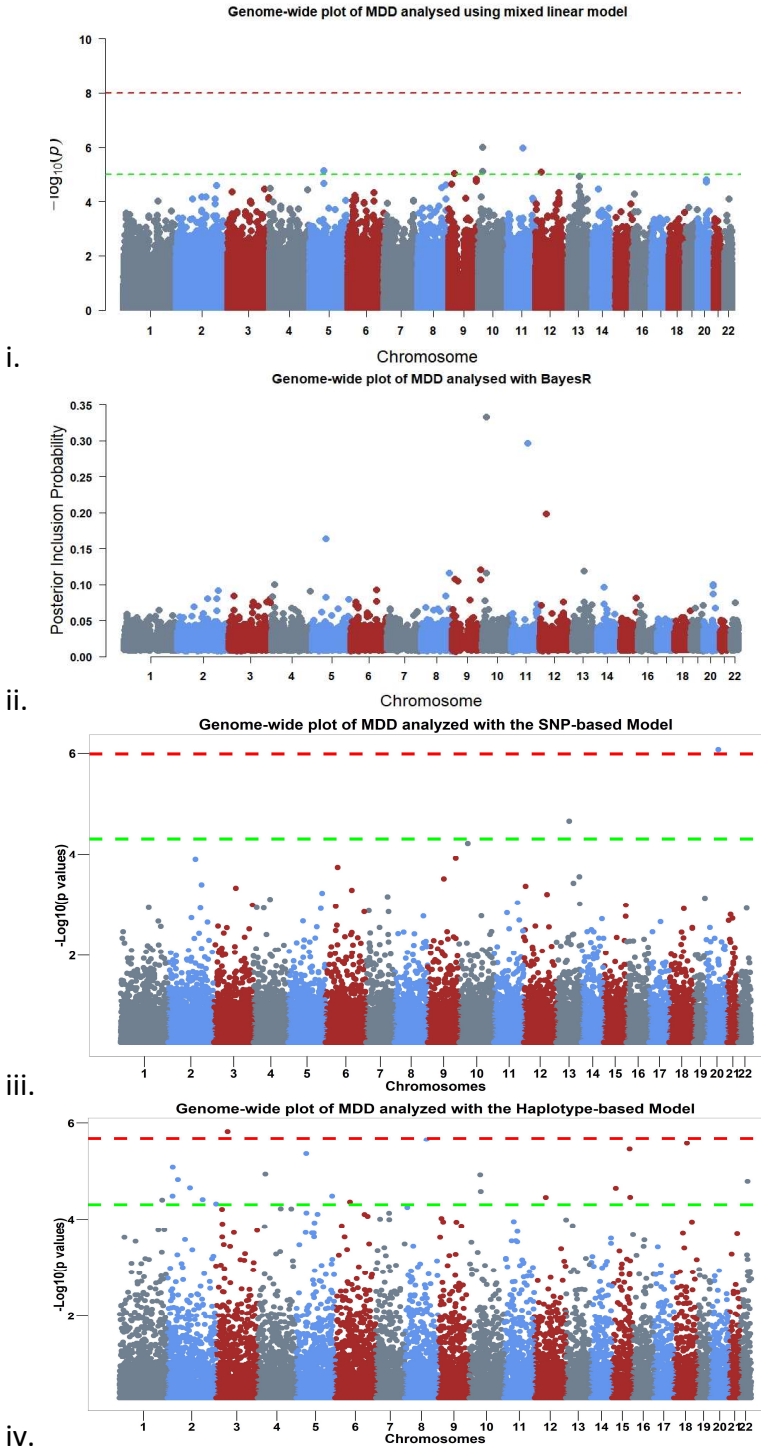
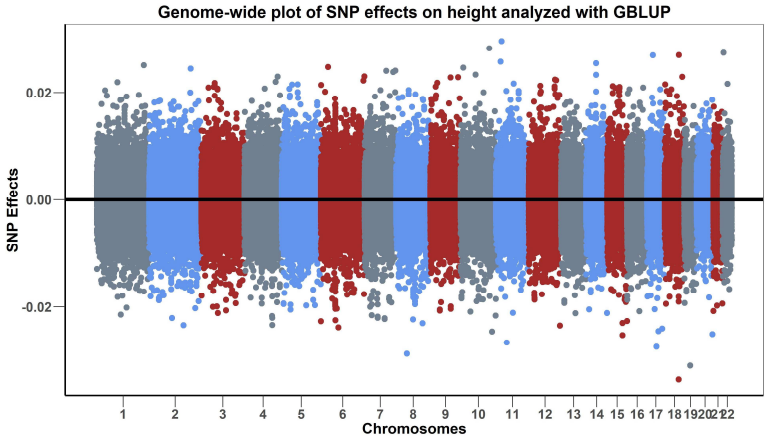



Figure 6.2. The genome-wide evidence of haplotype block association for Major Depressive Disorder. Analysis done with i. mixed linear model (MLM) ii. BayesR iii. SNP-based regional GREML model and iv. Haplotype-based regional GREML model. The points on the BayesR plot are plots of the posterior inclusion probability of SNPs. The points on the remaining plots are plots of -log10 of the p-values of SNP tested for MLM analysis and regions tested for the regional GREML analyses. The red dashed lines are the Bonferroni-corrected genome-wide significance threshold and the green dashed lines are suggestive significance threshold calculated to be p-value $< 5 \times 10^{-5}$. The association hits are more enriched as one moves from the SNP-based analyses to the haplotype-based analysis. This can be down to the fact that the genetic variation in MDD may be driven by rare genetic variants which may be in LD with haplotypes and thus their effects are captured by haplotype-based models.

The GBLUP results for both traits are shown in Figure 6.3. The assumption of GBLUP is that the effect of each genetic variant is sampled from the same normal distribution. Thus, the SNP effect smooths across the genome. The effect sizes of genetic variants in height are two orders of magnitude bigger than MDD. Height is highly heritable and is controlled by a lot of common genetic variants as opposed to MDD that has a low heritability and controlled by rare variants and this can explain the results observed.

i.




ii.

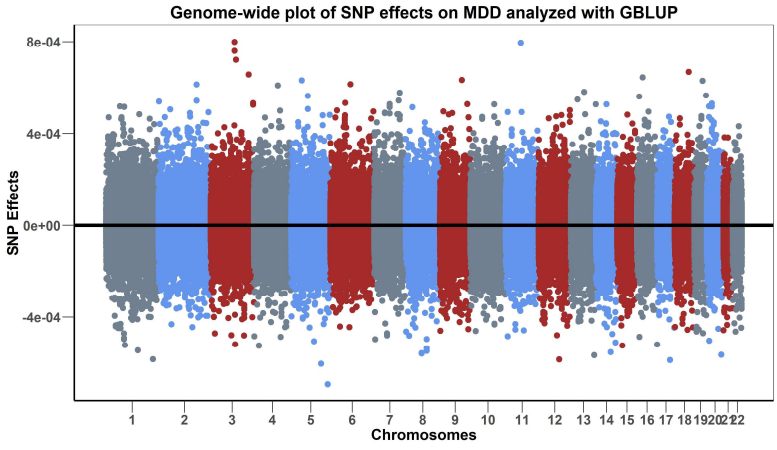



Figure 6.3. The genome-wide plot of SNP effects for height and Major Depressive Disorder. i. height ii. MDD. The effect sizes of genetic variants in height are two orders of magnitude bigger than MDD.

### 6.3.2 Comparison with published GWAS

For both traits, the SNPs in the regions that were significant at both genome-wide and suggestive p-values were compared to SNPs reported in the GWAS catalogue to be significant for the two traits. The GWAS catalogue was accessed on the 16th of July 2018. The results are presented in Table 6.2. The two models, SNP-based and haplotype-based models, taking all SNPs within significant blocks, identified 1,403 and 58 SNPs respectively for height, and 87 and 663 SNPs respectively for MDD.

Table 6.2. Comparison of SNPs within significant regions identified by both models and published GWAS results for height and MDD. The columns are the name of the trait, number of SNPs in regions identified by SNP-based (Sbm) and haplotype-based model (Hbm) with p-value $< 5 \times 10^{-5}$ and SNPs in published GWAS (pGWAS) for the traits, and the number of SNPs overlapping between the three.

| | Number of SNPs | | | Number of overlapping SNPs | | |
|---|---|---|---|---|---|---|
| **Trait** | **Sbm** | **Hbm** | **pGWAS** | **Sbm & Hbm** | **Sbm & pGWAS** | **Hbm & pGWAS** |
| **Height** | 1403 | 58 | 931 | 0 | 100 | 0 |
| **MDD** | 87 | 663 | 521 | 0 | 0 | 1 |

### 6.3.3 Association test for SNPs in the genome-wide significant region identified by Hbm for MDD

A SNP-based association test was performed for the region identified by the haplotype-based model for MDD in GS: SFHS to be genome-wide significant, p-value $< 2.04 \times 10^{-6}$. The results are shown in Table 6.3. Five SNPs within this region are nominally significant at p-value $< 0.05$. Four out of these five SNPs confer about 2% increased risk of the disease each. These four SNPs lie within the MYRIP gene sequence (Figure 6.4). The *MYRIP* gene is expressed in the brain (Ganat et al., 2012). SNPs in the *MYRIP* gene have been reported to be associated with brain processing

speed in the Lothian Birth cohort (Luciano et al., 2011), sleep duration (Gottlieb et al., 2007) and regulation of insulin levels (Waselle et al., 2003).

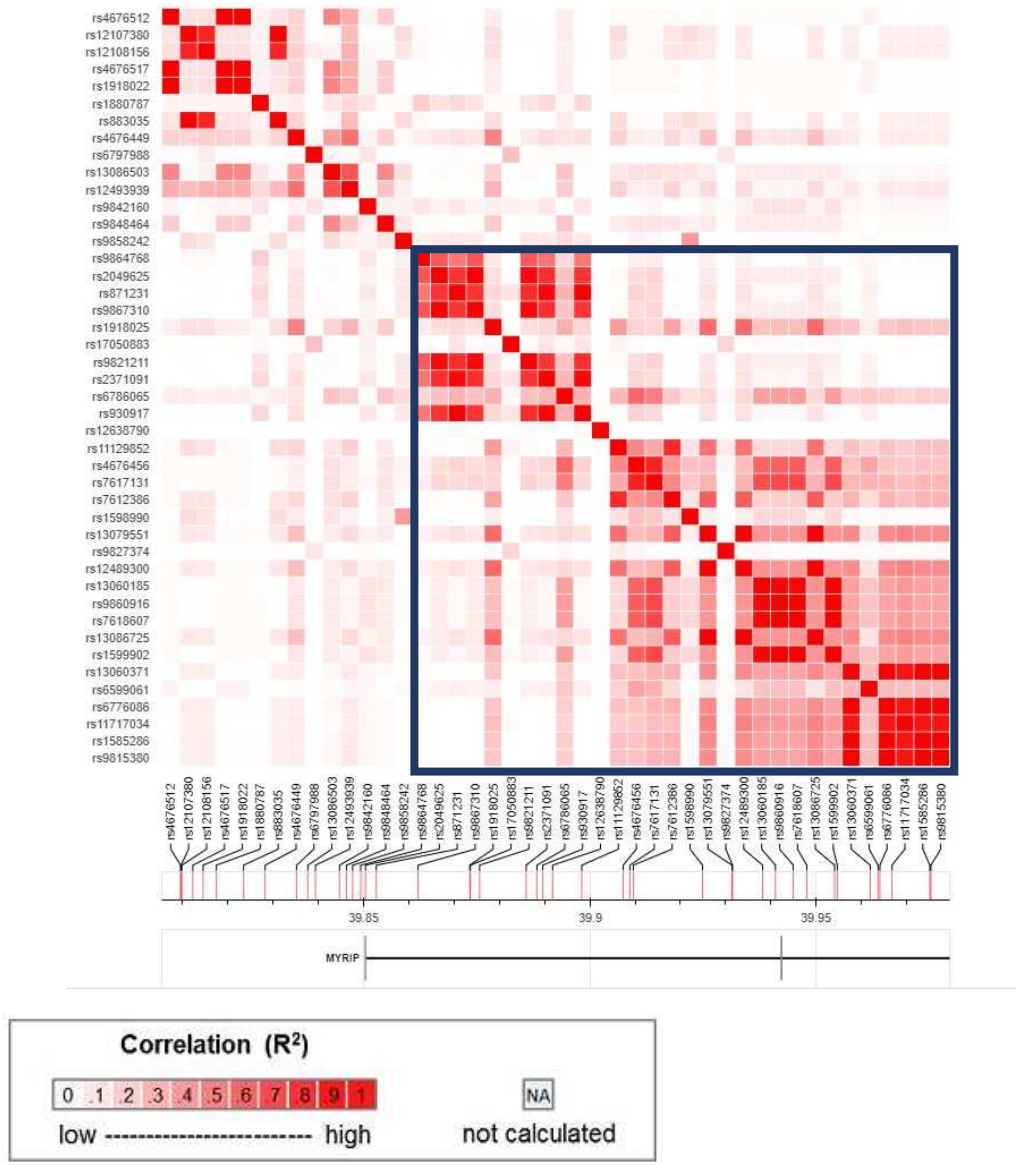



Figure 6.4. LD plot of most significant region for MDD identified by the haplotype-based model in the GS: SFHS cohort. The SNPs within this region lie in the MYRIP gene region. The SNPs in the MYRIP gene sequence (bounded by blue square) are in relatively high LD. The MYRIP gene has reported roles in processes that can be linked to MDD.

Table 6.3. SNP-based association test for regions identified by Hbm for MDD to be genome-wide significant. The columns are the regions, SNP ID, chromosome, genome position of SNP, minor allele frequency, odds ratio, a log of the odds ratio, standard error of log odds ratio and association p-value.

| SNP Information | | | | Depression Association | | | |
|---|---|---|---|---|---|---|---|
| **SNP ID** | **Chr** | **Pos** | **MAF** | **OR** | **Log (OR)** | **SE (logOR)** | **p** |
| **rs9842160** | 3 | 39844703 | 0.14 | 0.97 | -0.030 | 0.013 | 0.02 |
| **rs9858242** | 3 | 39847606 | 0.19 | 1.02 | 0.025 | 0.011 | 0.03 |
| **rs1599902** | 3 | 39954674 | 0.41 | 1.02 | 0.019 | 0.009 | 0.04 |
| **rs7618607** | 3 | 39947936 | 0.41 | 1.02 | 0.019 | 0.009 | 0.04 |
| **rs9860916** | 3 | 39944942 | 0.41 | 1.02 | 0.019 | 0.009 | 0.04 |

### 6.3.4 Replication of GS: SFHS MDD regions identified by Hbm in UK Biobank

Replication of the genomic regions that were identified as associated with MDD within the GS: SFHS by the haplotype-based model with p-value $< 5 \times 10^{-5}$ was sought in the UK Biobank Leeds cohort. The results are shown in Table 6.4. One region located on chromosome 15 was replicated with p-value $< 0.00416$.

## 6.4 Discussion

In this study, I have used a regional GREML method to perform a genome-wide analysis of height and MDD to identify genomic regions associated with height and MDD. This regional GREML method differs from a previous method (Cebamanos et al., 2014) in two ways. First is by the way regions are defined: genomic regions in my analysis are not defined by a fixed number of SNPs but are naturally defined using recombination hotspots from a reference human genome (Genome Reference Consortium Human Build 37 (International Human Genome Sequencing Consortium, 2004)). Second is that this method fits a haplotype-based GRM in a haplotype-based regional GREML model (Hbm).

Table 6.4. The replication of the genomic regions identified by the haplotype-based model to be associated with the GS: SFHS MDD phenotype at p-value $< 5 \times 10^{-5}$ in UK Biobank. The columns are chromosome number, region size in mega base-pairs, the number of SNPs typed in the region for the GS: SFHS cohort, the association p-values for GS: SFHS cohort, the number of SNPs typed in the region for the UK Biobank cohort and the association p-values for the UK Biobank cohort.

| Haplotype block | | GS: SFHS | | UK Biobank | |
|---|---|---|---|---|---|
| **Chr** | **Block size (Mb)** | **No. SNPs** | **P-value** | **No. SNPs** | **P-value** |
| **3** | 0.174935 | 44 | 1.50E-06 | 34 | 0.5 |
| **8** | 0.168933 | 40 | 2.21E-06 | 39 | 0.5 |
| **18** | 0.282808 | 49 | 2.66E-06 | 56 | 0.279 |
| **15** | 0.006625 | 6 | 3.50E-06 | 5 | 0.00416 |
| **5** | 0.006451 | 4 | 4.34E-06 | 6 | 0.466 |
| **2** | 0.025499 | 6 | 8.26E-06 | 10 | 0.5 |
| **4** | 0.222993 | 39 | 1.15E-05 | 49 | 0.484 |
| **10** | 0.046023 | 13 | 1.20E-05 | 13 | 0.5 |
| **2** | 0.118657 | 45 | 1.50E-05 | 33 | 0.126 |
| **22** | 0.474166 | 100 | 1.62E-05 | 132 | - |
| **2** | 0.220003 | 32 | 2.31E-05 | 20 | 0.461 |
| **15** | 0.028543 | 11 | 2.36E-05 | 15 | 0.349 |
| **10** | 0.049176 | 19 | 2.75E-05 | 15 | 0.5 |
| **5** | 0.018872 | 9 | 3.37E-05 | 7 | - |
| **2** | 0.026776 | 3 | 3.38E-05 | 6 | 0.5 |
| **15** | 0.042833 | 8 | 3.58E-05 | 12 | 0.0801 |
| **12** | 0.28385 | 46 | 3.61E-05 | 46 | 0.5 |
| **2** | 0.044811 | 12 | 3.99E-05 | 14 | 0.347 |
| **1** | 0.327998 | 104 | 4.08E-05 | 85 | - |
| **6** | 0.301042 | 24 | 4.46E-05 | 39 | 0.434 |
| **2** | 0.338596 | 49 | 4.87E-05 | 67 | - |

This study again differs from the haplotype association of MDD described by Howard et al. (2017) in two ways. That study set fixed haplotype windows sizes and used a sliding window approach to move along the genome to define haplotypes. And after the haplotypes are defined, they used a typical multiple regression GWAS to test for haplotype association with the MDD phenotype whereas I used the haplotypes in a regional GREML analysis. I draw comparisons with a model that fits a

conventional SNP-based GRM: a SNP-based regional GREML model (Sbm). I also compared the results obtained with the two models to those obtained with a linear mixed effect model (MLM), GBLUP and BayesR.

The SNP-based regional GREML analysis identified 31 genomic regions in GS: SFHS cohort associated with height with p-value $< 5 \times 10^{-5}$ out of which 14 were genome-wide (GW) significant, p-value $< 1.02 \times 10^{-6}$. For the MDD phenotype from the GS: SFHS cohort, the Sbm identified two associated regions with p-value $< 5 \times 10^{-5}$ of which one was GW significant. These regions harbour 1403 SNPs for height and 87 SNPs for MDD. Height is a highly polygenic trait with common genetic variants accounting for the majority of the additive genetic variation (Yang et al., 2015). The SNP-based regional GREML model, therefore, is better suited to capture SNP loci in height compared to MDD which is believed to driven by rare genetic variants. One hundred of the SNPs identified for height by the Sbm had been reported by other studies to be associated with height.

The SNP-based regional GREML analysis was able to capture a region on chromosome 10 that had been picked up by BayesR and MLM as being significantly associated with MDD. The SNP driving the association was found to lie in the *CACNB2* gene. This gene has been reported to be associated with major depressive disorder (Cross-Disorder Group of the Psychiatric Genomics Consortium, 2013).

Three genomic regions were identified for height by the haplotype-based regional GREML model with p-value $< 5 \times 10^{-5}$ none of which were genome-wide significant whereas, for MDD, the haplotype-based regional GREML model identified

21 regions of which one was genome-wide significant. Within these regions identified, there are 58 SNPs for height and 663 SNPs for MDD. The haplotype-based model works well for MDD because MDD is believed to be driven by rare genetic variants and the model can capture rare genetic variants. The haplotype model can capture rare variants because of the LD between rare variants (both typed and untyped) and the flanking markers which aggregate to form the haplotypes within genomic regions.

There were no overlaps between regions identified by the haplotype-based model and SNP-based models (Sbm, MLM, GBLUP and BayesR) for each of the two traits, which again support the hypothesis that the two classes of models complement each other in mapping associated loci.

The most significant region identified by the haplotype-based model for MDD is in the region of the *MYRIP* gene. Five SNPs within this region are individually significantly associated with MDD at the nominal level. Four of these SNPs lie within gene sequence of *MYRIP* and they each confer 2% disease risk. These nominally associated SNPs would have been missed by a conventional GWAS analysis because they will not reach genome-wide (GW) significance; although if the sample size were to be increased by say 10 times, then it might be possible to detect these SNPs at genome-wide significance level in a GWAS. Analysing the SNPs within the region as haplotypes, therefore, gave a greater power to detect the effect of the region at GW significance level with my relatively smaller sample size. This is because haplotypes can capture the joint effects of closely linked causal variants. Also, the haplotype-based analysis reduces the number of association tests performed; all the regions

that contained just one SNP were skipped in the analysis. This reduction in tests performed makes it possible to apply a less stringent Bonferroni-corrected threshold for genome-wide significance. The haplotype-based model is, therefore, a powerful approach to map low effect loci in the presence of low sample size.

The MYRIP gene identified in the GW significant region for MDD in the GS: SFHS cohort is expressed in the brain (Ganat et al., 2012). A SNP in this gene was reported to be associated with brain processing speed in the Lothian birth cohort (Luciano et al., 2011). Brain processing speed is an important cognitive function that is compromised in psychiatric illness such as schizophrenia and depression, and in old age. Also, a SNP in the MYRIP gene region associated with sleep duration (Gottlieb et al., 2007). Sleep duration outside the normal range (both short sleep and long sleep) has been found to be significantly associated with increased risk of depression (Mohan et al., 2017; Roberts and Duong, 2014; Watson et al., 2014; Zhai et al., 2015). The *MYRIP* gene is also reported to have a role in insulin secretion (Waselle et al., 2003) and low insulin levels have been linked to depression (Greenwood et al., 2015; Pearson et al., 2010; Webb et al., 2017).

I tried to replicate the 21 regions identified by the Hbm for GS: SFHS MDD phenotype in the UK Biobank cohort, but none of the regions was found to be genome-wide significant. Only one region on chromosome 15 was replicated with a p-value $< 0.00416$. The failure to replicate the regions in the UK Biobank cohort can be attributed to several things. First is that different SNPs may have been typed in the regions for the two cohorts which may generate different haplotypes, which can make replication of results impossible. Another reason is that the phenotype

ascertainment between the two cohorts was different: MDD cases in the GS: SFHS cohort was determined by a structured clinical interview for the diagnosis of mood disorders after participants had answered yes to one or more of initial screening questions whereas MDD cases in the UK Biobank were ascertained by participants answering yes to broad questions about having seen a GP or a psychiatrist for depression and related phenotypes. This difference in phenotype ascertainment introduces heterogeneity across the cohorts. And thus, the assayed regions may have different effects on the phenotype in each cohort. Another possibility is that the results obtained in the GS: SFHS cohort may be false positives, which will mean I was chasing shadows in the UK Biobank. However, the region identified by Hbm to be GW significant for MDD in the GS: SFHS cohort contained a gene which has SNPs reported to be associated with processes that can be linked to depression. This prior evidence increases support for the notion that the GS: SFHS results are real effects, but this is not conclusive. Again, maybe there was limited power to replicate the GS: SFHS results in the UK Biobank Leeds cohort.

In conclusion, I have shown for the two phenotypes analysed for the GS: SFHS cohort that the haplotype-based regional GREML model picks regions of the genome that explain genetic variance that is missed by the conventional SNP-based models. The haplotype approach analyses combinations of alleles instead of genotypes, and thus has greater power to detect regions for which the causative variant(s) is well tagged by a given haplotype or a few haplotypes. The haplotype method identified a novel risk region for MDD in the GS: SFHS cohort. This discovery was supported by a SNP-based associations test of the region that showed that four SNPs within the

region each had a risk allele that increases the risk of disease by 2% above the background risk. These SNPs lie in the *MYRIP* gene which has been shown to play a role in processes that affect depression. The results were not replicated in a UK Biobank subpopulation cohort possibly because of phenotype heterogeneity and different haplotype lengths within regions.

# Chapter 7

## 7 General Discussion

The development of efficient ways to estimate the genetic contribution to the phenotypic variation is critical to the genetic study of complex traits. This is because biased or inaccurate estimates will undermine the efforts aimed at unravelling the genetic contribution to the phenotypic variation. Therefore, in this thesis, my aim was to focus on the models used in the genome-wide association (GWA) study of complex human traits; investigating the performance of existing models and developing new methods to estimate genetic variance and map genomic loci associated with complex traits. Consequently, I have shown that normalisation of phenotype data helps minimise violations of GWA model assumptions and improves genetic associations. I have also presented the case for the use of a Bayesian mixture model in the GWA analysis of complex traits and shown that this Bayesian model does better than GBLUP in capturing genomic loci with effect especially for traits that are driven by a few large genomic loci. I then went on to show that utilizing haplotypes in a regional GREML setting can uncover components of the genetic variance that are missed by conventional GWAS.

In this final chapter, I revisit the main findings of this PhD project and situate them in the wider research context: discussing the relevance of these findings and shedding light on some of the limitations of the study. I also make some suggestions for future work.

To investigate the impact of violation of GWA analysis model assumptions, I used a linear mixed model to perform a genome-wide association analysis of concentrations of eight urine electrolytes measured in 2,934 participants from the Generation Scotland: Scottish Family Health Study. The association analysis was performed on the residuals obtained after regressing normal-transformed and untransformed trait values on covariates. The traits were transformed prior to correcting for the covariates and not after, as normalizing corrected residuals can reintroduce the linear relationship between traits and covariates (Pain et al., 2018) and increase type I errors. The results showed that normal transformation improved evidence of SNP associations in most traits. The reason for this can be that transformation minimizes the violation of model assumptions which then improves the results from the analyses by minimizing the chances of committing either Type I or II errors (Osborne, 2010). The magnitude of the effect sizes of SNPs in the transformed data were, however, smaller compared to the untransformed data. This was the case because by transforming data, the nature of the relationship between the phenotypes and the genetic variants changes which can change the effect sizes in either direction, up or down as was in this case. I found that the significantly associated SNPs for these traits lie within regions of the genome that have genes nearby some of which have reported roles in processes or phenotypes relating to

kidney disease and function. It would have been great to explore these genes to understand their roles in the context of kidney disease development, but this was beyond the remit of this research and also because the heritability estimates from these traits were small to modest. There have been very scant GWAS reports on these traits which makes it difficult to assess my findings, but that makes these findings uniquely important, in terms of expanding the GWAS catalogue.

The simulation study in chapter three suggests that BayesR (Erbe et al., 2012) is a good analytical tool for studying the genetic architecture of complex traits. This was further confirmed in chapter four where I applied the BayesR model to the urine phenotypes from the GS: SFHS. The model was able to unearth the underlying genetic architecture of these traits. The model has been shown to work not only in continuous traits but also in binary traits (Moser et al., 2015). The method gives an estimate of the number of loci affecting a trait and the proportion of the additive genetic variance they explain which potentially becomes useful to the efforts of trying to dissect the genetic architecture of complex traits. Using a simulation study, I explored the BayesR model in detail and drew comparisons with the GBLUP model (VanRaden, 2008). The results showed that the two models, BayesR and GBLUP give good estimates of the heritability of traits. The BayesR model, however, was better than GBLUP in capturing simulated effect SNPs. The GBLUP model assumes a normal distribution with a common variance for sampling the allelic substitutions effect for all marker loci. So, for any given locus, the GBLUP model does not have any prior on how big the locus effect is. The model does not differentiate between the marker loci: the fact that some may have a large effect, and some may be explaining a small

effect. This is because the GBLUP model primarily assumes traits are controlled by polygenes, i.e. many genes with each gene having a small effect on the trait. What happens then if a trait is affected by major genes? Because there is evidence from linkage studies that some human complex traits may be mainly driven by major genes with large effects; examples are given in some obesity-related traits (Comuzzie et al., 1997; Hager et al., 1998), type 2 diabetes (Duggirala et al., 1999; Grant et al., 2006; Reynisdottir et al., 2003) and BMI (Luke et al., 2003; Moslehi et al., 2003). Which is why one goes towards a Bayesian model that allows the modelling of genetic variants with large effect. So, instead of assuming that all loci are expected to explain the same variance, one assumes that traits should have for example 90% of the loci explaining a small variance and 10% explaining a large variance. Based on this assumption then one can start to generalise and suggest that the distribution of the variances associated with marker loci is maybe an inverse chi-square as in Bayes B (Meuwissen et al., 2001), so there is a small proportion of marker loci that explains a large effect and most of the loci explain only very small effects. What is being done here is really just exploring the distribution of the SNP effects. That is instead of having a normal distribution for sampling SNP effects (as in GBLUP), other distributions that allow a much greater probability that some markers will have a much larger effect than others are used. In so doing, you are changing the possibility of finding SNPs with large effects. The Bayesian mixture model used in this thesis, BayesR, is built on the same idea. The model allows one to model the effect of SNPs as coming from four normal distributions instead of one: i.e. model a group of loci from a distribution with no effect on the trait, then another group that individually

have a small effect from a distribution with a small variance, then another group from a distribution with moderate variance and then another group still with bigger variance and so on. The BayesR model will, therefore, be better at modelling a genetic architecture with loci of different effect sizes than GBLUP as was shown in the simulation study in chapter three. The BayesR model is also well suited for dissecting the genetic architecture of complex traits. This is because the model allows the estimation of the number of loci affecting trait variation, their contribution to the additive genetic variance, and give the distribution of allelic effects.

The accuracy of prediction obtained in chapter three, however, would suggest that GBLUP performs better at predicting phenotypes than BayesR. This result differs from those reported by Moser et al. (2015) who reported comparatively higher BayesR accuracies than GBLUP. The reason for this can be ascribed to the close family relationship structures in my dataset. About a third of the individuals used in my analysis had close relatives in the dataset, thus the random assignment of individuals to the training and validation set makes it possible that some of the individuals in the testing set will be related to individuals in the training set. The accuracy of prediction depends on the SNPs that are in LD with causal loci and on SNPs that capture the relationship structure between the individuals in the training and validation dataset (Habier and Fernando, 2013; Habier et al., 2008). Therefore, the availability of lots of closely related individuals in the dataset is bound to improve prediction accuracy if the model can capture these relationships. Between the two models, GBLUP is more effective in capturing genetic relationships because it fits all markers into the prediction model whereas BayesR fits markers sampled to have an

effect on the trait. So, in the case of GBLUP, the SNPs that don't contribute anything to the trait variance will still contribute towards estimating the relationships and thus will contribute something towards prediction. So, in datasets with lots of close relationships, GBLUP will mostly perform better than BayesR. Bermingham et al. (2015) similarly reported slightly higher accuracies for GBLUP than a Bayesian model (Bayes C) in participants of the Orkney Complex Disease Study (ORCADES) (McQuillan et al., 2008). If pairwise relationships were very distant, GBLUP will never do that well because the model smooths SNP effects across the genome. Whereas for the BayesR approach, the model picks out SNPs that really have effects and predicts based on those SNPs. So, for BayesR, relationships between individuals don't really matter as long as the effect SNPs that explain a large enough proportion of the genetic variance are identified and used in the prediction. The BayesR model on average picked between 3,000 and 7,000 SNPs to have an effect on simulated traits. Although in reality it correctly mapped on average between 50 and 200 of the 5,520 SNPs that were simulated to have an effect on the traits. The 20 large effect loci were however effectively mapped in the traits with moderate and higher heritability. The prediction accuracies obtained with the BayesR model will be largely driven by these large effect loci. In livestock populations, every individual is related to every other (Groeneveld et al., 2010) or at least everyone has a relative in the data, and thus GBLUP works very well. But the GBLUP model may not work very well going to human populations where relationships are less because it requires related individuals at some level. Whereas a Bayesian model potentially may have an advantage there because it

Investigating the genetic control of complex traits identifies SNP effects and puts weights on big SNP effects and does prediction with them.

The BayesR accuracies obtained in chapter three were between 41% and 46% of the theoretical maximum attainable for these simulated traits which is good for this size of training dataset (about 2,000 individuals). So, the argument again would be to get more data. This is because when you get a lot of data, the training set is then more powerful at differentiating between the models because you've got more information which can be incorporated into the prediction. Fitting a normal distribution, a Laplace distribution, an inverse chi-square distribution or four normal distributions etc. would not do very much. Because by the time you get enough data to estimate SNP effects quite reliably, you are not doing very much regression back on the marker effects, i.e. you have got enough data, you are getting quite confident in your estimation of SNP effect and there isn't much regression back towards the mean. Therefore, the influence of the prior distribution gets smaller when you have big data. So, yes, some analytical methods will have better power than others in capturing variants associated with complex traits because as I have shown different models are needed to accommodate different trait architectures but increasing the sample size will continue to be the best strategy in the effort to dissect the genetic architecture of complex traits. That is why I increased the data size to include about 20,000 individuals in the analysis in the next chapters of the thesis.

Following the Bayesian analysis of complex traits, I moved on to perform a regional GREML analysis that analysed the genome in haplotype blocks bounded by recombination hotspots. This analysis method draws strongly on the work of Shirali

et al. (2018) who performed a similar analysis on other human traits. This study differs from the work of Shirali et al. (2018) by the way phenotypes are simulated, which expands on the work done previously. The method is grounded on the hypothesis that analysing haplotypes can capture portions of the genetic variance that may be missed by conventional SNP-based analysis. Thus, the haplotype-based method will be complementary to existing GWA analysis methods. I confirmed this hypothesis in a simulation study that analysed 20 replicates of two types of phenotypes in which SNPs are simulated to have an effect and three types of phenotypes in which haplotypes are simulated to have an effect.

The haplotype-based method, however, struggled to accurately capture the marker effect in regions with very long haplotypes. I tried to resolve this by breaking up these long haplotypes, but this resulted in reduced test statistics. I, therefore, recommend the use of a lower recombination rate threshold instead of artificially breaking up the haplotypes in such regions with long haplotypes to alleviate this issue. One other possibility to consider for this haplotype method is that the documented recombination hotspots from the Reference Consortium Human Build 37 (International Human Genome Sequencing Consortium, 2004) that are used may not reflect the actual recombination hotspots in the population analysed. Therefore, it is possible that the hotspot boundaries used are inaccurate which then can negatively impact on the method and make it less effective especially in the analysis of real phenotypes. This was, however, found not to be a major problem when the method was implemented in chapter six to analyse real phenotype data from GS: SFHS. The haplotype boundaries defined in the GS: SFHS cohort using the

recombination hotspots from the reference genome worked fine and a novel locus was mapped for MDD. This locus harbours a gene which has reported functions in processes that are linked to MDD.

In chapter six of the thesis, I discussed the main arguments that concern the issue of genome-wide association (GWA) studies of MDD not being successful in identifying genetic variants that are associated with MDD at genome-wide significance level. I then presented the argument that GWA studies of MDD that utilizes haplotypes can offer some solution to the problem. Although a couple of studies (Howard et al., 2017; Zeng et al., 2017a) have highlighted the importance of using haplotypes in GWA study of MDD, there has not been a strong focus on employing haplotypes in MDD GWA research. As such, I used chapter six to provide further insights into the use of haplotypes as one of the ways forward in uncovering associated genetic variants for MDD. The haplotype-based regional GREML analysis of MDD phenotype from the GS: SFHS cohort identified a novel risk region for MDD. Four SNPs within this region were found to confer about 2% risk of the disease each in a SNP-based associations test. These SNPs were found to lie in the sequence of the *MYRIP* gene which has reported roles in processes that affect depression. However, I failed to replicate the results in the UK Biobank Leeds cohort. One reason for this can be differences in phenotype ascertainment which introduces heterogeneity across the two cohorts. Another reason can be that the two cohorts have different LD structures arising from different recombination hotspots which possibly means they have different haplotype structures. The second reason may not be strong enough considering the two cohorts share the same ancestral population. Perhaps

another possibility for the failure to replicate the finding is that it may be a false positive. But in any case, one way forward is to perform a haplotype-based regional GREML analysis of the full genome of the full UK Biobank cohort. After which a meta-analysis of significant regions between the UK Biobank and GS: SFHS cohorts can be performed. Meta-analysis of MDD has been shown to work by Wray et al. (2018) who identified 44 independent loci significantly associated with MDD in a GWA meta-analysis of MDD. This, therefore, warrants a further investigation of the method through a meta-analysis of regions between the two cohorts.

## 7.1 Conclusion and future considerations

In conclusion, the genome-wide association (GWA) studies have provided a great deal of insight into questions that bear on the genetic control of complex traits. This thesis has explored the underlying assumptions of some of the analytical approaches used in the GWA analysis of complex traits and offered novel insights into how to incorporate other genetic variants like haplotypes in the GWA study of complex traits.

There are additional things to consider beyond what is presented in this thesis to support the effort to disentangle the genetic underpinning of complex traits. One key thing to be considered for future research is an investigation of how much information is lost when phenotype data is transformed. Pain et al. (2018) have already led this effort using simulation and real data to explore what happens to the phenotype – covariates relationship when phenotype data is transformed. There is still some work to be done in terms of investigating how transformation impacts

effect sizes, and whether it can increase false positives and so on. One way to do this is, perhaps, to simulate phenotypes with SNP effect sizes sampled from a non-normal distribution and investigate how much information is lost through normalisation. In addition to this, one can investigate the effect of normalisation on simulated phenotypes in which the environmental effects are sampled from a non-normal distribution. There are several possibilities in terms of the type of distribution the effects are sampled from and also what type of effect to simulate; whether additive, dominant or recessive, espistatic or a combination of all four.

For the BayesR model, fitting a mixture of four normal distributions for sampling the effects of SNPs on all traits can be deemed overly simplistic. Thus, further research that tries to work out the best way to fit an ideal number of distributions for different types of traits will improve upon the model.

Additionally, the haplotype regional GREML method in its current state, although shown to be useful in this thesis and by others (Shirali et al., 2018), still needs further improvement to make its use more appealing. It is currently computationally intensive which can make it particularly painful to use. I attempted to apply the method to the full set of participants in the UK Biobank cohort but that was fraught with several computing challenges. The most challenging limiting factors were RAM and CPU requirements. The regional GREML model involves the use of GRMs that get larger with increasing sample size. The RAM requirement for calculating these matrices increases by a factor of $N^2$ where $N$ is the sample size. So, when the sample size doubles, the RAM requirement for calculating the GRMs quadruples and this immediately became a problem for me. Because moving from

my sample size of about 20,000 individuals in the GS: SFHS to the UK Biobank sample size of 500,000 individuals, increased $N$ by 25 times. This increased the RAM requirement by 625 times. Similarly, the computational power requirements for the GREML analysis of one region grow by $N^3$. So, again the number of processor cores required to analyse one region in the full UK Biobank cohort increased by more than 15,625 times compared to the GS: SFHS data. And I had over 30,000 genomic regions to analyse, which brings additional problems of compute cost in pounds sterling and also compute time, which basically meant I couldn't do it in the time scale of my PhD. This is because on average, the GS: SFHS analyses required about an hour and a half of computational time to analyse one genomic region on four CPU cores with 8Gib of RAM per core (32Gib RAM total). This essentially means that for 30,000 genomic regions, the total compute time required was about 45,000 CPU hours per model (two models) for each phenotype (two phenotypes). And scaling this up to analyse the whole of the UK Biobank data will require at least 15,625 CPU cores with more than 20,000Gib of RAM in total, which simply becomes intractable for a PhD. So, although I have argued that it is good to have a large sample size because you get a significant boost in power to detect effects, fine mapping efforts like the haplotype regional GREML analysis can easily become intractable in large samples. The method has been shown in this thesis to be good for mapping novel genetic loci. However, there is not yet a practical way of implementing this kind of analysis on such large cohorts as the UK Biobank. It is still possible all the same, just that it will require highly skilled computer programmers with proficient knowledge in high performance computing structure and parallelisation to make this possible. Therefore,

collaborative research efforts with people in computer science and engineering should be pushed in the future.

Finally, another thing that should be worthy of consideration in future research efforts into this method is to investigate how other types of data such as exome sequence and whole genome sequence data can be incorporated in the analysis. With the plummeting costs of whole genome resequencing, research focus in GWA studies is already shifting towards sequence data analysis. Although whole-genome sequence data analysis would allow the incorporation of all the genetic variants that drive the phenotypic variation, there may still be some variants whose individual effects may be too small to be picked up in a conventional GWA analysis. However, regionally analysing sequence information can help overcome this because multiple small effect variants in a region can add up to a substantial regional effect which can be captured by a regional SNP GRM or tagged by a haplotype. Haplotypes can tag sequence information well and thus this method which can utilize haplotype information offers a lot of promise in the coming days.

“Without data, you're just another person with an opinion.”

— W. Edwards Deming

# References

Almasy, L., and Blangero, J. (2001). Endophenotypes as quantitative risk factors for psychiatric disease: rationale and study design. Am. J. Med. Genet. *105*, 42–44.

Antonarakis, S.E., and Beckmann, J.S. (2006). Mendelian disorders deserve more attention. Nat. Rev. Genet. *7*, 277–282.

Aulchenko, Y.S. (2011). Chapter 9 - Effects of Population Structure in Genome-wide Association Studies. In Analysis of Complex Disease Association Studies, E.Z. Morris, ed. (San Diego: Academic Press), pp. 123–156.

Aulchenko, Y.S., Ripke, S., Isaacs, A., and Duijn, C.M. (2007). GenABEL: an R library for genome-wide association analysis. Bioinformatics *23*.

Balding, D.J. (2006). A tutorial on statistical methods for population association studies. Nat. Rev. Genet. *7*, 781–791.

Barrett, J.C. (2011). Chapter 2 - Population Genetics and Linkage Disequilibrium. In Analysis of Complex Disease Association Studies, E.Z. Morris, ed. (San Diego: Academic Press), pp. 15–23.

Beavis, W.D. (1997). QTL Analyses: Power, Precision, and Accuracy. In Molecular Dissection of Complex Traits, A.H. Paterson, ed. (Boca Raton, FL: CRC Press), pp. 145–161.

Benjamin, D., Cesarini, D., Loos, M., Dawes, C., Koellinger, P., Magnusson, P., Chabris, C., Conley, D., Laibson, D., Johannesson, M., et al. (2012). The Genetic Architecture of Economic and Political Preferences. Proc. Natl. Acad. Sci. *109*, 8026–8031.

Bermingham, M.L., Pong-Wong, R., Spiliopoulou, A., Hayward, C., Rudan, I., Campbell, H., Wright, A.F., Wilson, J.F., Agakov, F., Navarro, P., et al. (2015). Application of high-dimensional feature selection: evaluation for genomic prediction in man. Sci. Rep. *5*, 10312.

Björkegren, J.L.M., Kovacic, J.C., Dudley, J.T., and Schadt, E.E. (2015). Genome-Wide Significant Loci: How Important Are They?: Systems Genetics to Understand Heritability of Coronary Artery Disease and Other Common Complex Disorders. J. Am. Coll. Cardiol. *65*, 830–845.

Bodmer, W., and Bonilla, C. (2008). Common and rare variants in multifactorial susceptibility to common diseases. Nat. Genet. *40*, 695–701.

Botstein, D., White, R.L., Skolnick, M., and Davis, R.W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am. J. Hum. Genet. *32*, 314–331.

Box, G.E.P., and Cox, D.R. (1964). An Analysis of Transformations. J. R. Stat. Soc. Ser. B Methodol. *26*, 211–252.

Burton, P.R., Clayton, D.G., Cardon, L.R., Craddock, N., Deloukas, P., Duncanson, A., Kwiatkowski, D.P., McCarthy, M.I., Ouwehand, W.H., Samani, N.J., et al. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature *447*, 661–678.

Bush, W.S., and Moore, J.H. (2012). Chapter 11: Genome-Wide Association Studies. PLoS Comput Biol *8*, e1002822.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2017). Genome-wide genetic data on ~500,000 UK Biobank participants. BioRxiv 166298.

de los Campos, G., Gianola, D., and Allison, D.B. (2010). Predicting genetic predisposition in humans: the promise of whole-genome markers. Nat. Rev. Genet. *11*, 880–886.

Canela-Xandri, O., Law, A., Gray, A., Woolliams, J.A., and Tenesa, A. (2015). A new tool called DISSECT for analysing large genomic data sets using a Big Data approach. Nat. Commun. *6*, ncomms10162.

Cebamanos, L., Gray, A., Stewart, I., and Tenesa, A. (2014). Regional Heritability Advanced Complex Trait Analysis for GPU and Traditional Parallel Architectures. Bioinformatics btt754.

Chambers, J.C., Zhang, W., Lord, G.M., van der Harst, P., Lawlor, D.A., Sehmi, J.S., Gale, D.P., Wass, M.N., Ahmadi, K.R., Bakker, S.J.L., et al. (2010). Genetic loci influencing kidney function and chronic kidney disease. Nat. Genet. *42*, 373–375.

Chen, W.-M., and Abecasis, G.R. (2007). Family-based association tests for genomewide association scans. Am. J. Hum. Genet. *81*, 913–926.

Chen, X., Kuja-Halkola, R., Rahman, I., Arpegård, J., Viktorin, A., Karlsson, R., Hägg, S., Svensson, P., Pedersen, N.L., and Magnusson, P.K.E. (2015). Dominant Genetic Variation and Missing Heritability for Human Complex Traits: Insights from Twin versus Genome-wide Common SNP Models. Am. J. Hum. Genet. *97*, 708–714.

Chial, H. (2008). Rare Genetic Disorders: Learning About Genetic Disease Through Gene Mapping, SNPs, and Microarray Data. Nat. Educ. *1*, 192.

Cirulli, E.T., and Goldstein, D.B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat. Rev. Genet. *11*, 415–425.

Clarke, A.J., and Cooper, D.N. (2010). GWAS: heritability missing in action? Eur. J. Hum. Genet. *18*, 859–861.

Collins, F.S., Green, E.D., Guttmacher, A.E., and Guyer, M.S. (2003). A vision for the future of genomics research. Nature *422*, 835–847.

Comuzzie, A.G., Hixson, J.E., Almasy, L., Mitchell, B.D., Mahaney, M.C., Dyer, T.D., Stern, M.P., MacCluer, J.W., and Blangero, J. (1997). A major quantitative trait locus determining serum leptin levels and fat mass is located on human chromosome 2. Nat. Genet. *15*, 273–276.

Consortium, T. 1000 G.P. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.

Cross-Disorder Group of the Psychiatric Genomics Consortium (2013). Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. Lancet Lond. Engl. *381*, 1371–1379.

Cross-Disorder Group of the Psychiatric Genomics Consortium, Lee, S.H., Ripke, S., Neale, B.M., Faraone, S.V., Purcell, S.M., Perlis, R.H., Mowry, B.J., Thapar, A., Goddard, M.E., et al. (2013). Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. Nat. Genet. *45*, 984–994.

Daetwyler, H.D., Villanueva, B., and Woolliams, J.A. (2008). Accuracy of Predicting the Genetic Risk of Disease Using a Genome-Wide Approach. PLoS ONE *3*, e3395.

Daetwyler, H.D., Pong-Wong, R., Villanueva, B., and A Woolliams, J. (2010). The Impact of Genetic Architecture on Genome-Wide Evaluation Methods. Genetics *185*, 1021–1031.

Daly, M.J., Rioux, J.D., Schaffner, S.F., Hudson, T.J., and Lander, E.S. (2001). High-resolution haplotype structure in the human genome. Nat. Genet. *29*, 229–232.

Darwin, C. (1859). On the Origin of Species by Means of Natural Selection, Or, The Preservation of Favoured Races in the Struggle for Life (J. Murray).

Delaneau, O., Marchini, J., and Zagury, J.-F. (2012). A linear complexity phasing method for thousands of genomes. Nat. Methods *9*, 179–181.

Delaneau, O., Zagury, J.-F., and Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. Nat. Methods *10*, 5–6.

Duggirala, R., Blangero, J., Almasy, L., Dyer, T.D., Williams, K.L., Leach, R.J., O'Connell, P., and Stern, M.P. (1999). Linkage of type 2 diabetes mellitus and of age at onset to a genetic location on chromosome 10q in Mexican Americans. Am. J. Hum. Genet. *64*, 1127–1140.

Erbe, M., Hayes, B.J., Matukumalli, L.K., Goswami, S., Bowman, P.J., Reich, C.M., Mason, B.A., and Goddard, M.E. (2012). Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J. Dairy Sci. *95*, 4114–4129.

Falconer, D.S. (1960). Introduction to quantitative genetics (New York,: Ronald Press Co).

Fernando, R.L., and Grossman, M. (1989). Marker assisted selection using best linear unbiased prediction. Genet. Sel. Evol. *21*, 467.

First, M.B., Spitzer, R.L., Gibbon, M., and Williams, J.B.W. (2002). Structured Clinical Interview for DSM-IV-TR Axis I Disorders, Research Version, Non-patient Edition.

Fisher, R.A. (1918). The Correlation between Relatives on the Supposition of Mendelian Inheritance. R Soc Edinb. Trans *52*, 399–433.

Forte, L.R., Fan, X., and Hamra, F.K. (1996). Salt and water homeostasis: uroguanylin is a circulating peptide hormone with natriuretic activity. Am. J. Kidney Dis. Off. J. Natl. Kidney Found. *28*, 296–304.

Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P., Leal, S.M., et al. (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851–861.

Freedman, D.A. (2009). Statistical Models: Theory and Practice (Cambridge University Press).

Freedman, M.L., Reich, D., Penney, K.L., McDonald, G.J., Mignault, A.A., and Patterson, N. (2004). Assessing the impact of population stratification on genetic association studies. Nat Genet *36*.

Ganat, Y.M., Calder, E.L., Kriks, S., Nelander, J., Tu, E.Y., Jia, F., Battista, D., Harrison, N., Parmar, M., Tomishima, M.J., et al. (2012). Identification of embryonic stem cell-derived midbrain dopaminergic neurons for engraftment. J. Clin. Invest. *122*, 2928–2939.

Ghosh, S., and Collins, F.S. (1996). The geneticist's approach to complex disease. Annu. Rev. Med. *47*, 333–353.

Gianola, D. (2013). Priors in Whole-Genome Regression: The Bayesian Alphabet Returns. Genetics *194*, 573–596.

Gianola, D., Campos, G. de los, Hill, W.G., Manfredi, E., and Fernando, R. (2009). Additive Genetic Variability and the Bayesian Alphabet. Genetics *183*, 347–363.

Gibbs, R.A., Belmont, J.W., Hardenbol, P., Willis, T.D., Yu, F., Yang, H., Ch'ang, L.-Y., Huang, W., Liu, B., Shen, Y., et al. (2003). The International HapMap Project. Nature *426*, 789–796.

Gibson, G. (2010). Hints of hidden heritability in GWAS. Nat. Genet. *42*, 558–560.

Gonzalez-Recio, O., Daetwyler, H.D., MacLeod, I.M., Pryce, J.E., Bowman, P.J., Hayes, B.J., and Goddard, M.E. (2015). Rare Variants in Transcript and Potential Regulatory Regions Explain a Small Percentage of the Missing Heritability of Complex Traits in Cattle. PLoS ONE *10*, e0143945.

Gottlieb, D.J., O'Connor, G.T., and Wilk, J.B. (2007). Genome-wide association of sleep and circadian phenotypes. BMC Med. Genet. *8*, S9.

Grant, S.F.A., Thorleifsson, G., Reynisdottir, I., Benediktsson, R., Manolescu, A., Sainz, J., Helgason, A., Stefansson, H., Emilsson, V., Helgadottir, A., et al. (2006). Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. Nat. Genet. *38*, 320–323.

Greenwood, E.A., Pasch, L.A., Shinkai, K., Cedars, M.I., and Huddleston, H.G. (2015). Putative role for insulin resistance in depression risk in polycystic ovary syndrome. Fertil. Steril. *104*, 707-714.e1.

Groeneveld, L.F., Lenstra, J.A., Eding, H., Toro, M.A., Scherf, B., Pilling, D., Negrini, R., Finlay, E.K., Jianlin, H., Groeneveld, E., et al. (2010). Genetic diversity in farm animals--a review. Anim. Genet. *41 Suppl 1*, 6–31.

Gudbjartsson, D.F., Helgason, H., Gudjonsson, S.A., Zink, F., Oddson, A., Gylfason, A., Besenbacher, S., Magnusson, G., Halldorsson, B.V., Hjartarson, E., et al. (2015). Large-scale whole-genome sequencing of the Icelandic population. Nat. Genet. *47*, 435–444.

Guo, H., Gao, P., Li, Y., Zhu, D., and Wang, J. (2007). [Association between uroguanylin G-247A polymorphism and blood pressure/fluid and electrolytes homeostasis]. Zhonghua Xin Xue Guan Bing Za Zhi *35*, 233–236.

Habier, D., and Fernando, R.L. (2013). Genomic BLUP Decoded: A Look into the Black Box of Genomic Prediction. Genetics *194*.

Habier, D., Fernando, R., and Dekkers, J. (2008). The Impact of Genetic Relationship Information on Genome-Assisted Breeding Values. Genetics *177*, 2389–2397.

Hager, J., Dina, C., Francke, S., Dubois, S., Houari, M., Vatin, V., Vaillant, E., Lorentz, N., Basdevant, A., Clement, K., et al. (1998). A genome-wide scan for human obesity genes reveals a major susceptibility locus on chromosome 10. Nat. Genet. *20*, 304–308.

Hall, M.-H., and Smoller, J.W. (2010). A New Role for Endophenotypes in the GWAS Era: Functional Characterization of Risk Variants. Harv. Rev. Psychiatry *18*, 67–74.

Harris, H., and Hopkinson, D.A. (1976). Handbook of Enzyme Electrophoresis in Human Genetics (Amsterdam: North-Holland Publishing Company).

Haseman, J.K., and Elston, R.C. (1972). The investigation of linkage between a quantitative trait and a marker locus. Behav. Genet. *2*, 3–19.

Hayes, B.J., Pryce, J., Chamberlain, A.J., Bowman, P., and Goddard, M.E. (2010). Genetic Architecture of Complex Traits and Accuracy of Genomic Prediction: Coat Colour, Milk-Fat Percentage, and Type in Holstein Cattle as Contrasting Model Traits. PLoS Genet. *6*, e1001139.

Held, L., and Bové, D.S. (2014). Bayesian Inference. In Applied Statistical Inference, (Springer Berlin Heidelberg), pp. 167–219.

Hemani, G., Knott, S., and Haley, C. (2013). An Evolutionary Perspective on Epistasis and the Missing Heritability. PLoS Genet *9*, e1003295.

Hill, W.G. (1984). Quantitative Genetics Part I: Explanation And Analysis Of Continuous Variation (Van Nostrand Reinhold).

Hishida, A., Nakatochi, M., Akiyama, M., Kamatani, Y., Nishiyama, T., Ito, H., Oze, I., Nishida, Y., Hara, M., Takashima, N., et al. (2018). Genome-Wide Association Study of Renal Function Traits: Results from the Japan Multi-Institutional Collaborative Cohort Study. Am. J. Nephrol. *47*, 304–316.

Hoggart, C.J., Whittaker, J.C., Iorio, M.D., and Balding, D.J. (2008). Simultaneous Analysis of All SNPs in Genome-Wide and Re-Sequencing Association Studies. PLOS Genet. *4*, e1000130.

Howard, D.M., Hall, L.S., Hafferty, J.D., Zeng, Y., Adams, M.J., Clarke, T.-K., Porteous, D.J., Nagy, R., Hayward, C., Smith, B.H., et al. (2017). Genome-wide haplotype-based association analysis of major depressive disorder in Generation Scotland and UK Biobank. Transl. Psychiatry *7*.

Howard, D.M., Adams, M.J., Shirali, M., Clarke, T.-K., Marioni, R.E., Davies, G., Coleman, J.R.I., Alloza, C., Shen, X., Barbu, M.C., et al. (2018). Genome-wide association study of depression phenotypes in UK Biobank identifies variants in excitatory synaptic pathways. Nat. Commun. *9*, 1470.

International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. Nature *431*, 931–945.

Kendler, K.S., Neale, M.C., Kessler, R.C., Heath, A.C., and Eaves, L.J. (1993). A test of the equal-environment assumption in twin studies of psychiatric illness. Behav. Genet. *23*, 21–27.

Kennedy, M.A. (2001). Mendelian Genetic Disorders. In ELS, (John Wiley & Sons, Ltd), p.

Kiel, D.P., Demissie, S., Dupuis, J., Lunetta, K.L., Murabito, J.M., and Karasik, D. (2007). Genome-wide association with bone mass and geometry in the Framingham Heart Study. BMC Med. Genet. *8 Suppl 1*, S14.

Kimura, M. (1983). The Neutral Theory of Molecular Evolution. Camb. Univ. Press N. Y.

Kinoshita, H., Fujimoto, S., Nakazato, M., Yokota, N., Date, Y., Yamaguchi, H., Hisanaga, S., and Eto, T. (1997). Urine and plasma levels of uroguanylin and its molecular forms in renal diseases. Kidney Int. *52*, 1028–1034.

Köttgen, A., Pattaro, C., Böger, C.A., Fuchsberger, C., Olden, M., Glazer, N.L., Parsa, A., Gao, X., Yang, Q., Smith, A.V., et al. (2010). Multiple New Loci Associated with Kidney Function and Chronic Kidney Disease: The CKDGen consortium. Nat. Genet. *42*, 376–384.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Lee, N.P.Y., Tong, M.K., Leung, P.P., Chan, V.W., Leung, S., Tam, P.-C., Chan, K.-W., Lee, K.-F., Yeung, W.S.B., and Luk, J.M. (2006). Kidney claudin-19: localization in distal tubules and collecting ducts and dysregulation in polycystic renal disease. FEBS Lett. *580*, 923–931.

Lee, S.H., Yang, J., Chen, G.-B., Ripke, S., Stahl, E.A., Hultman, C.M., Sklar, P., Visscher, P.M., Sullivan, P.F., Goddard, M.E., et al. (2013). Estimation of SNP Heritability from Dense Genotype Data. Am. J. Hum. Genet. *93*, 1151–1155.

Levinson, D.F., Mostafavi, S., Milaneschi, Y., Rivera, M., Ripke, S., Wray, N.R., and Sullivan, P.F. (2014). Genetic studies of major depressive disorder: Why are there no GWAS findings, and what can we do about it? Biol. Psychiatry *76*, 510–512.

Lin, Y.-J., Liao, W.-L., Wang, C.-H., Tsai, L.-P., Tang, C.-H., Chen, C.-H., Wu, J.-Y., Liang, W.-M., Hsieh, A.-R., Cheng, C.-F., et al. (2017). Association of human height-related genetic variants with familial short stature in Han Chinese in Taiwan. Sci. Rep. *7*.

Loewe, L. (2008). Negative selection. Nat. Educ. *1*, 59.

Lubke, G.H., Hottenga, J.J., Walters, R., Laurin, C., de Geus, E.J.C., Willemsen, G., Smit, J.H., Middeldorp, C.M., Penninx, B.W.J.H., Vink, J.M., et al. (2012). Estimating the genetic variance of major depressive disorder due to all single nucleotide polymorphisms. Biol. Psychiatry *72*, 707–709.

Luciano, M., Hansell, N.K., Lahti, J., Davies, G., Medland, S.E., Räikkönen, K., Tenesa, A., Widen, E., McGhee, K.A., Palotie, A., et al. (2011). Whole genome association

scan for genetic polymorphisms influencing information processing speed. Biol. Psychol. *86*, 193–202.

Luke, A., Wu, X., Zhu, X., Kan, D., Su, Y., and Cooper, R. (2003). Linkage for BMI at 3q27 region confirmed in an African-American population. Diabetes *52*, 1284–1287.

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., et al. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. *45*, D896–D901.

Maher, B. (2008). Personal genomes: The case of the missing heritability. Nat. News *456*, 18–21.

Major Depressive Disorder Working Group of the Psychiatric GWAS Consortium, Ripke, S., Wray, N.R., Lewis, C.M., Hamilton, S.P., Weissman, M.M., Breen, G., Byrne, E.M., Blackwood, D.H.R., Boomsma, D.I., et al. (2013). A mega-analysis of genome-wide association studies for major depressive disorder. Mol. Psychiatry *18*, 497–511.

Makowsky, R., Pajewski, N.M., Klimentidis, Y.C., Vazquez, A.I., Duarte, C.W., Allison, D.B., and de los Campos, G. (2011). Beyond Missing Heritability: Prediction of Complex Traits. PLoS Genet *7*, e1002051.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature *461*, 747–753.

Marchini, J. (2015). UK Biobank Phasing and Imputation Documentation.

Marin, J.-M., Mengersen, K., and Robert, C.P. (2005). Bayesian modelling and inference on mixtures of distributions. Handb. Stat. *25*, 459–507.

McCarroll, S.A. (2008). Extending genome-wide association studies to copy-number variation. Hum. Mol. Genet. *17*, R135–R142.

McCarthy, M.I., Abecasis, G.R., Cardon, L.R., Goldstein, D.B., Little, J., Ioannidis, J.P.A., and Hirschhorn, J.N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat. Rev. Genet. *9*, 356–369.

McQuillan, R., Leutenegger, A.-. L., Abdel-Rahman, R., Franklin, C.S., Pericic, M., and Barac-Lauc, L. (2008). Runs of homozygosity in European populations. Am J Hum Genet *83*.

Meuwissen, T.H.E., Hayes, B.J., and Goddard, M.E. (2001). Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics *157*, 1819–1829.

Mohan, J., Xiaofan, G., and Yingxian, S. (2017). Association between sleep time and depression: a cross-sectional study from countries in rural Northeastern China. J. Int. Med. Res. *45*, 984–992.

Moore, R. (2001). The “Rediscovery” of Mendel’s Work. Bioscene *27*, 13–23.

Morgan, T.H., and Bridges, C.B. (1916). Sex-Linked Inheritance in Drosophila (The Carnegie Institution of Washington).

Moser, G., Lee, S.H., Hayes, B.J., Goddard, M.E., Wray, N.R., and Visscher, P.M. (2015). Simultaneous Discovery, Estimation and Prediction Analysis of Complex Traits Using a Bayesian Mixture Model. PLOS Genet. *11*, e1004969.

Moslehi, R., Goldstein, A.M., Beerman, M., Goldin, L., Bergen, A.W., and Framingham Heart Study (2003). A genome-wide linkage scan for body mass index on Framingham Heart Study families. BMC Genet. *4 Suppl 1*, S97.

Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., and Erlich, H. (1986). Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. Cold Spring Harb. Symp. Quant. Biol. *51 Pt 1*, 263–273.

Nagamine, Y., Pong-Wong, R., Navarro, P., Vitart, V., Hayward, C., Rudan, I., Campbell, H., Wilson, J., Wild, S., Hicks, A.A., et al. (2012). Localising Loci underlying Complex Trait Variation Using Regional Genomic Relationship Mapping. PLoS ONE *7*, e46501.

Nicoletti, P., Cartsos, V.M., Palaska, P.K., Shen, Y., Floratos, A., and Zavras, A.I. (2012). Genomewide pharmacogenetics of bisphosphonate-induced osteonecrosis of the jaw: the role of RBMS3. The Oncologist *17*, 279–287.

Osborne, J.W. (2010). Improving Your Data Transformations: Applying the Box-Cox Transformation. Pract. Assess. Res. Eval. *15*.

Pai, A.A., Pritchard, J.K., and Gilad, Y. (2015). The Genetic and Mechanistic Basis for Variation in Gene Regulation. PLOS Genet. *11*, e1004857.

Pain, O., Dudbridge, F., and Ronald, A. (2018). Are your covariates under control? How normalization can re-introduce covariate effects. Eur. J. Hum. Genet. *26*, 1194–1201.

Park, J.-H., Wacholder, S., Gail, M.H., Peters, U., Jacobs, K.B., Chanock, S.J., and Chatterjee, N. (2010). Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nat. Genet. *42*, 570–575.

Pearson, S., Schmidt, M., Patton, G., Dwyer, T., Blizzard, L., Otahal, P., and Venn, A. (2010). Depression and Insulin Resistance. Diabetes Care *33*, 1128–1133.

Pe'er, I., de Bakker, P.I.W., Maller, J., Yelensky, R., Altshuler, D., and Daly, M.J. (2006). Evaluating and improving power in whole-genome association studies using fixed marker sets. Nat. Genet. *38*, 663–667.

Pe'er, I., Yelensky, R., Altshuler, D., and Daly, M.J. (2008). Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. Genet. Epidemiol. *32*, 381–385.

Picard, F. (2007). An introduction to mixture models. Stat. Syst. Biol. Res. Rep.

Piras, D., Zoledziewska, M., Cucca, F., and Pani, A. (2017). Genome-Wide Analysis Studies and Chronic Kidney Disease. Kidney Dis. *3*, 106–110.

Powell, J.E., Visscher, P.M., and Goddard, M.E. (2010). Reconciling the analysis of IBD and IBS in complex trait studies. Nat. Rev. Genet. *11*, 800–805.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. *38*, 904–909.

Pritchard, J.K. (2001). Are rare variants responsible for susceptibility to complex diseases? Am. J. Hum. Genet. *69*, 124–137.

Pruim, R.J., Welch, R.P., Sanna, S., Teslovich, T.M., Chines, P.S., Gliedt, T.P., Boehnke, M., Abecasis, G.R., and Willer, C.J. (2010). LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics *26*, 2336–2337.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., et al. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am. J. Hum. Genet. *81*, 559–575.

Rahbi, H., Narayan, H., Jones, D.J.L., and Ng, L.L. (2012). The uroguanylin system and human disease. Clin. Sci. *123*, 659–668.

Reddi, A.S. (2014). Fluid, Electrolyte and Acid-Base Disorders Clinical Evaluation and Management.

Reich, D.E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P.C., Richter, D.J., Lavery, T., Kouyoumjian, R., Farhadian, S.F., Ward, R., et al. (2001). Linkage disequilibrium in the human genome. Nature *411*, 199–204.

Reynisdottir, I., Thorleifsson, G., Benediktsson, R., Sigurdsson, G., Emilsson, V., Einarsdottir, A.S., Hjorleifsdottir, E.E., Orlygsdottir, G.T., Bjornsdottir, G.T., Saemundsdottir, J., et al. (2003). Localization of a susceptibility gene for type 2 diabetes to chromosome 5q34-q35.2. Am. J. Hum. Genet. *73*, 323–335.

Reynolds, C.A., and Finkel, D. (2015). A Meta-analysis of Heritability of Cognitive Aging: Minding the “Missing Heritability” Gap. Neuropsychol. Rev. *25*, 97–112.

Roberts, R.E., and Duong, H.T. (2014). The Prospective Association between Sleep Deprivation and Depression among Adolescents. Sleep *37*, 239–244.

Robinson, G.K. (1991). That BLUP is a Good Thing: The Estimation of Random Effects. Stat. Sci. *6*, 15–32.

Sabeti, P.C., Reich, D.E., Higgins, J.M., Levine, H.Z.P., Richter, D.J., Schaffner, S.F., Gabriel, S.B., Platko, J.V., Patterson, N.J., McDonald, G.J., et al. (2002). Detecting recent positive selection in the human genome from haplotype structure. Nature *419*, 832–837.

Seng, K.C., and Seng, C.K. (2008). The success of the genome-wide association approach: a brief story of a long struggle. Eur. J. Hum. Genet. *16*, 554–564.

Sham, P.C., and Purcell, S.M. (2014). Statistical power and significance testing in large-scale genetic studies. Nat. Rev. Genet. *15*, 335–346.

Shirali, M., Knott, S.A., Pong-Wong, R., Navarro, P., and Haley, C.S. (2018). Haplotype Heritability Mapping Method Uncovers Missing Heritability of Complex Traits. Sci. Rep. *8*, 4982.

Shriner, D., Vaughan, L.K., Padilla, M.A., and Tiwari, H.K. (2007). Problems with Genome-Wide Association Studies. Science *316*, 1840–1842.

Slatkin, M. (2009). Epigenetic Inheritance and the Missing Heritability Problem. Genetics *182*, 845–850.

Smith, B.H., Campbell, H., Blackwood, D., Connell, J., Connor, M., and Deary, I.J. (2006). Generation Scotland: the Scottish Family Health Study; a new resource for researching genes and heritability. BMC Med Genet *7*.

Smith, B.H., Campbell, A., Linksted, P., Fitzpatrick, B., Jackson, C., and Kerr, S.M. (2012). Cohort profile: Generation Scotland: Scottish Family Health Study (GS:SFHS). The study, its participants and their potential for genetic research on health and illness. Int J Epidemiol *42*.

Soldatov, N.M. (2015). CACNB2: An Emerging Pharmacological Target for Hypertension, Heart Failure, Arrhythmia and Mental Disorders. Curr. Mol. Pharmacol. *8*, 32–42.

Southern, E.M. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. J. Mol. Biol. *98*, 503–517.

Speed, D., and Balding, D.J. (2015). Relatedness in the post-genomic era: is it still useful? Nat. Rev. Genet. *16*, 33–44.

Speed, D., Hemani, G., Johnson, M.R., and Balding, D.J. (2012). Improved Heritability Estimation from Genome-wide SNPs. Am. J. Hum. Genet. *91*, 1011–1021.

Stephens, M., and Balding, D.J. (2009). Bayesian statistical methods for genetic association studies. Nat. Rev. Genet. *10*, 681–690.

Strandén, I., and Garrick, D.J. (2009). Technical note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. J. Dairy Sci. *92*, 2971–2975.

Stranger, B.E., Stahl, E.A., and Raj, T. (2011). Progress and Promise of Genome-Wide Association Studies for Human Complex Trait Genetics. Genetics *187*, 367–383.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLOS Med. *12*, e1001779.

Thomas, D. (2010). Gene-Environment-Wide Association Studies: Emerging Approaches. Nat. Rev. Genet. *11*, 259–272.

Thompson, E.A., and Shaw, R.G. (1990). Pedigree analysis for quantitative traits: variance components without matrix inversion. Biometrics *46*, 399–413.

Tibshirani, R. (1994). Regression Shrinkage and Selection Via the Lasso. J. R. Stat. Soc. Ser. B *58*, 267–288.

VanRaden, P.M. (2008). Efficient methods to compute genomic predictions. J. Dairy Sci. *91*, 4414–4423.

Visscher, P.M., Brown, M.A., McCarthy, M.I., and Yang, J. (2012). Five years of GWAS discovery. Am. J. Hum. Genet. *90*, 7–24.

Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., and Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. Am. J. Hum. Genet. *101*, 5–22.

Vivante, A., Kleppa, M.-J., Schulz, J., Kohl, S., Sharma, A., Chen, J., Shril, S., Hwang, D.-Y., Weiss, A.-C., Kaminski, M.M., et al. (2015). Mutations in TBX18 Cause Dominant Urinary Tract Malformations via Transcriptional Dysregulation of Ureter Development. Am. J. Hum. Genet. *97*, 291–301.

Vormfelde, S.V., and Brockmöller, J. (2007). On the value of haplotype-based genotype-phenotype analysis and on data transformation in pharmacogenetics and -genomics. Nat. Rev. Genet. *8*.

Wain, L.V., Shrine, N., Miller, S., Jackson, V.E., Ntalla, I., Soler Artigas, M., Billington, C.K., Kheirallah, A.K., Allen, R., Cook, J.P., et al. (2015). Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank. Lancet Respir. Med. *3*, 769–781.

Wallace, C., Newhouse, S.J., Braund, P., Zhang, F., Tobin, M., Falchi, M., Ahmadi, K., Dobson, R.J., Marçano, A.C.B., Hajat, C., et al. (2008). Genome-wide Association Study Identifies Genes for Biomarkers of Cardiovascular Disease: Serum Urate and Dyslipidemia. Am. J. Hum. Genet. *82*, 139–149.

Wang, B., Sverdlov, S., and Thompson, E. (2017). Efficient Estimation of Realized Kinship from SNP Genotypes. Genetics genetics.116.197004.

Waselle, L., Coppola, T., Fukuda, M., Iezzi, M., El-Amraoui, A., Petit, C., and Regazzi, R. (2003). Involvement of the Rab27 Binding Protein Slac2c/MyRIP in Insulin Exocytosis. Mol. Biol. Cell *14*, 4103–4113.

Watson, N.F., Harden, K.P., Buchwald, D., Vitiello, M.V., Pack, A.I., Strachan, E., and Goldberg, J. (2014). Sleep Duration and Depressive Symptoms: A Gene-Environment Interaction. Sleep *37*, 351–358.

Webb, M., Davies, M., Ashra, N., Bodicoat, D., Brady, E., Webb, D., Moulton, C., Ismail, K., and Khunti, K. (2017). The association between depressive symptoms and insulin resistance, inflammation and adiposity in men and women. PLOS ONE *12*, e0187448.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., et al. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acid Res. *42*.

Wieczorek, S., Holle, J.U., Müller, S., Fricke, H., Gross, W.L., and Epplen, J.T. (2010). A functionally relevant IRF5 haplotype is associated with reduced risk to Wegener's granulomatosis. J. Mol. Med. Berl. Ger. *88*, 413–421.

Wray, N.R., and Maier, R. (2014). Genetic Basis of Complex Genetic Disease: The Contribution of Disease Heterogeneity to Missing Heritability. Curr. Epidemiol. Rep. *1*, 220–227.

Wray, N.R., Goddard, M.E., and Visscher, P.M. (2007). Prediction of individual genetic risk to disease from genome-wide association studies. Genome Res. *17*, 1520–1528.

Wray, N.R., Yang, J., Hayes, B.J., Price, A.L., Goddard, M.E., and Visscher, P.M. (2013). Pitfalls of predicting complex traits from SNPs. Nat. Rev. Genet. *14*, 507–515.

Wray, N.R., Ripke, S., Mattheisen, M., Trzaskowski, M., Byrne, E.M., Abdellaoui, A., Adams, M.J., Agerbo, E., Air, T.M., Andlauer, T.M.F., et al. (2018). Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. Nat. Genet. *50*, 668–681.

Wuttke, M., Wong, C.S., Wühl, E., Epting, D., Luo, L., Hoppmann, A., Doyon, A., Li, Y., CKDGen Consortium, Sözeri, B., et al. (2016). Genetic loci associated with renal function measures and chronic kidney disease in children: the Pediatric Investigation for Genetic Factors Linked with Renal Progression Consortium. Nephrol. Dial. Transplant. Off. Publ. Eur. Dial. Transpl. Assoc. - Eur. Ren. Assoc. *31*, 262–269.

Xia, C., Amador, C., Huffman, J., Trochet, H., Campbell, A., Porteous, D., Scotland, G., Hastie, N.D., Hayward, C., Vitart, V., et al. (2016). Pedigree- and SNP-Associated Genetics and Recent Environment are the Major Contributors to Anthropometric and Cardiometabolic Trait Variation. PLOS Genet. *12*, e1005804.

Xue, J.L., Ma, J.Z., Louis, T.A., and Collins, A.J. (2001). Forecast of the number of patients with end-stage renal disease in the United States to the year 2010. J. Am. Soc. Nephrol. JASN *12*, 2753–2758.

Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. *42*, 565–569.

Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: A Tool for Genome-wide Complex Trait Analysis. Am. J. Hum. Genet. *88*, 76–82.

Yang, J., Bakshi, A., Zhu, Z., Hemani, G., Vinkhuyzen, A.A.E., Lee, S.H., Robinson, M.R., Perry, J.R.B., Nolte, I.M., van Vliet-Ostaptchouk, J.V., et al. (2015). Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Nat. Genet. *47*, 1114–1120.

Yang, T.-L., Guo, Y., Li, J., Zhang, L., Shen, H., Li, S.M., Li, S.K., Tian, Q., Liu, Y.-J., Papasian, C.J., et al. (2013). Gene-gene interaction between RBMS3 and ZNF516 influences bone mineral density. J. Bone Miner. Res. Off. J. Am. Soc. Bone Miner. Res. *28*, 828–837.

Zaitlen, N., Kraft, P., Patterson, N., Pasaniuc, B., Bhatia, G., Pollack, S., and Price, A.L. (2013). Using Extended Genealogy to Estimate Components of Heritability for 23 Quantitative and Dichotomous Traits. PLoS Genet *9*, e1003520.

Zeng, Y., Navarro, P., Shirali, M., Howard, D.M., Adams, M.J., Hall, L.S., Clarke, T.-K., Thomson, P.A., Smith, B.H., Murray, A., et al. (2017a). Genome-wide Regional Heritability Mapping Identifies a Locus Within the TOX2 Gene Associated With Major Depressive Disorder. Biol. Psychiatry *82*, 312–321.

Zeng, Y., Navarro, P., Fernandez-Pujals, A.M., Hall, L.S., Clarke, T.-K., Thomson, P.A., Smith, B.H., Hocking, L.J., Padmanabhan, S., Hayward, C., et al. (2017b). A Combined Pathway and Regional Heritability Analysis Indicates NETRIN1 Pathway Is Associated With Major Depressive Disorder. Biol. Psychiatry *81*, 336–346.

Zhai, L., Zhang, H., and Zhang, D. (2015). SLEEP DURATION AND DEPRESSION AMONG ADULTS: A META-ANALYSIS OF PROSPECTIVE STUDIES. Depress. Anxiety *32*, 664–670.

Zhu, Z., Bakshi, A., Vinkhuyzen, A.A.E., Hemani, G., Lee, S.H., Nolte, I.M., van Vliet-Ostaptchouk, J.V., Snieder, H., Esko, T., Milani, L., et al. (2015). Dominance Genetic Variation Contributes Little to the Missing Heritability for Human Complex Traits. Am. J. Hum. Genet. *96*, 377–385.

Zuk, O., Hechter, E., Sunyaev, S.R., and Lander, E.S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. Proc. Natl. Acad. Sci. *109*, 1193–1198.